Contributions to the Current Debates on the Financial Markets and on Environmental Policy in China

Dissertation
submitted to the
Faculty of Business, Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktorin der Wirtschaftswissenschaften, Dr. oec.
(corresponds to Doctor of Philosophy, PhD)

presented by

Yingnan Zhao
from China

approved in September 2016 at the request of

Prof. Dr. Dr. Josef Falkinger
Prof. Dr. Fabrizio Zilibotti

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, 21.09.2016

Chairman of the Doctoral Board: Prof. Dr. Steven Ongena

Acknowledgements

I started as a fast-track PhD student at Josef Falkinger's Chair of Public Finance and Macroeconomics at the University of Zurich in 2011. Time flies! Throughout the past five years, I have benefited a lot from people around me and now it is time to say thank you.

I am most indebted to my supervisor Josef Falkinger, for his guidance, mentoring, patience and encouragement on my road to rigorous economic research. I am deeply inspired and influenced by his insights and passion for economics, and his pursuit for fundamentals and precision. I know this influence will benefit me for my entire life and I truly appreciate it.

I also wish to express my gratitude to my co-advisor Fabrizio Zilibotti. I thank him for his illuminating feedback and comments on my dissertation. I am very happy and honored to get my first job offer after my PhD studies from him. I am very much looking forward to more inspiring discussions with him.

Moreover, I am thankful to my co-author Sabrina Studer. She is a perfect model of Swiss quality. I enjoyed all the afternoons we spent on our joint work and benefited from her understanding and intuitions on economic principles.

I also want to thank the professors, post-docs and fellow PhD students in the Department of Economics, who have inspired me and helped me to improve my work during lectures and seminars. Furthermore, I am grateful to the current and former members of the chair for all their assistance.

I am grateful to my friends. Their company and support make my time at the University of Zurich enjoyable and unforgettable.

Finally, I wish to thank my parents for their unconditional love and infinite support that have accompanied me in each day of my life.

Zurich, June 2016 Yingnan Zhao

Contents

Part I

Dissertation Overview

Dissertation Overview

In the past decades, two topics have attracted public attention both inside and outside academia: Financial markets and environment. Environment is fundamental for the well being of human life and the financial sector plays a key role for the functioning of a modern economy, due to its principle interconnection with the real economy.

First, when governments and regulators discuss whether stricter regulations should be imposed on banks in the aftermath of the financial crisis in 2007-2008, two questions are fundamental: How is the availability of credit to firms influenced by aggregate fluctuations of the economy? And what do stricter regulations mean for the real sector?

Second, if we take a closer look at the financial sector, it is clear from data that tremendous structural change has taken place in the past thirty years: This is not only reflected by an increasing weight of the financial sector relative to the production sector, but also by a shift within the financial sector from conventional banking towards modern new finance (e.g., commodity contracts, securities, etc). A natural question is therefore: What drives such twofold structural change?

The first two chapters of this dissertation study the interconnection between the financial sector and the real economy from the above-mentioned two perspectives, respectively. The third chapter converts the focus to the other pressing issue: Environment.

China has experienced substantial economic growth, but also severe environmental deterioration in the last decades. It is both of academic interest and of the interest of fellow households to understand potential reasons that induce such situation. A series of observations point to a role of the political institutions in China: First, China's environmental pollution surpasses other countries in their similar phase of economic development and industrial structure. Second, local governments in China have low incentives to improve environment, due to career concerns and availability of fiscal budgets. And lastly, there are increasing cases of household protest due to environmental issues. These observations lay the ground for my third project.

The first paper *Bank lending and firm dynamics in general equilibrium* (jointly with Sabrina Studer) characterizes a long-term dynamic lending relationship between banks and firms in a general equilibrium framework. First, firms need bank loans to produce, and banks provide loans with deposits from workers. The productivity of firm production is a random variable, and the realized productivity in each period is private information

of firms. To avoid information asymmetry, profit-maximizing banks provide loans and ask for repayments through long-term dynamic contracts. Like in standard literature, the dynamic contract is recursively determined with promised value as state variable, where the promised value in the context of this paper is firms' discounted future income (output net of repayments to banks). Since the availability of bank loans determines firms' budget for employing input factors, the lending relationship between banks and firms determines the size of firms in the economy, as well as important characteristics of firm dynamics, such as firms' growth and the volatility of growth. In a second step, we embed the dynamic loan contract into a general equilibrium framework by endogenously determining the occupational choice of households (to become a worker or an entrepreneur). This allows us to get the share of entrepreneurs in equilibrium and the factor prices. The theoretical contribution of the paper is that we merge two strands of research: Lending relationships between banks and firms in general equilibrium, and impact of firm financing on firm dynamics (in a partial equilibrium). The quantitative results of the model are in line with empirical observations: Young firms are more financially constrained, and the situation is alleviated in repeated interactions with banks. In addition, young firms are characterized by faster growth but more volatility in growth. Furthermore, comparative statics analysis of the model allows us to assess the impact of an increase in productivity volatility, and of an increasing stringency in bank regulation (in terms of higher reserve ratio) on equilibrium factor prices, entrepreneurship and firm dynamics.

The second project *Explaining structural change towards and within the financial sector* (jointly with Josef Falkinger and Sabrina Studer) studies in a three-sector OLG framework potential drivers of the twofold structural change. The three sectors are a good sector that produces consumption and investment goods, and two financial sectors that provide financial services for transforming household savings into future consumption possibilities. High- and low-skilled labor are used in production of goods and services. And capital is additionally used in the good sector. Sectors differ in skill intensity. Households make saving and portfolio decisions to maximize lifetime expected utility. To transform savings in the safe and in risky assets, financial services from the two financial sectors are demanded. In general equilibrium, the skill premium, which is the relative wage of high- to low-skilled labor, and the sectoral structure of an economy are determined. Furthermore, we do comparative statics analysis with respect to fundamentals (i.e., households endowment, supply of skill and directed technical change). Using the wage premium as an indicator of inequality, we identify the following channels as common drivers that contribute simultaneously to two salient features of the recent development: The twofold structural change as well as the increasing inequality. Specifically, the channels are uniform productivity growth across sectors, biased technical change and increasing completeness of financial

market. We further extend the baseline model to allow for frictions, which in fact characterize financial market (e.g., fixed costs, rents or participation constraints in the financial sector, as well as distorted portfolio choice due to erroneous belief and set-up capital for firms), and discuss their impact on structural change and on inequality. In the end, we calibrate the model using US data from 1980-2014, and show that the potential drivers of change suggested by the theoretical analysis are consistent with the data.

The third project *Environmental policy and social stability in China* analyzes reasons of China's environmental deterioration from a political economy perspective. This paper provides a framework for analyzing "high output / high pollution" issue in China. It accounts for China's political system and addresses its key elements: The hierarchy system between the central government and local officials, the central government's cadre system, local officials' fiscal investment and increasing cases of household protest. The model includes a central government, local officials and households from N regions. The benevolent central government chooses the environmental policy that maximizes household welfare; however, it needs local officials to implement the policy by regional investments. Households act to maintain their utility (from consumption and environment) above an exogenous threshold: When their utility falls below, they protest against their local official. Local officials have career concerns (in pursuit of promotion and avoiding demotion due to local protests): They maximize their expected income by allocating local fiscal budget on production-related and environment-related infrastructure. Higher local production increases their probability of being promoted, whereas overly high production deteriorates environment and increases probability of household protest in the region. The tradeoff of the central government is between a moderate pollution abatement policy on the one side and a decreasing probability of policy success under more thorough environmental policy. We compare local officials' investment allocation under two scenarios: In an equilibrium where the local officials are incentivized by career concerns, and in a social optimum with benevolent local officials. We find that promotion incentives induce an overly high investment in production-related infrastructure; the extent of overinvestment depends on local officials' ability and the strength of environmental policy of the central government. As a consequence, households have higher consumption but worse environmental quality in the equilibrium with incentivized officials compared to the social optimum with benevolent officials. Additionally, the central government chooses a weaker environmental policy to avoid local officials' incentive to overinvest, which further lowers the equilibrium environmental quality below the social optimum one.

The structure of the dissertation is as follows: The three papers are presented in Part II and the respective appendices are provided in Part III. The bibliography is in Part IV and Part V includes my curriculum vitae.

Part II

Research Papers

1 Bank lending and firm dynamics in general equilibrium

Joint with Sabrina Studer

1.1 Introduction

Access to financing is one of the main issues firms are dealing with. In general, financial constraints determine firms' development and their size distribution (Angelini and Generale, 2008). Especially for small and medium-sized firms with constrained access to bond or equity markets (The Economist, 2015), bank loans account for the primary part of external financing (Berger and Udell, 2002). This project analyzes how entrepreneurs and banks interact by modeling long-term credit relationships between them. Long-term credit relationships help to overcome information asymmetries through dynamic contracting. To the best of our knowledge we are the first who deal with such a long-term lending relationship in a general equilibrium framework which allows us to determine endogenously both the share of entrepreneurs as well as important aspects of firm dynamics such as size, growth and variance of growth of firms.

A key point of our model is the assumption of information asymmetry. This is, entrepreneurs have private knowledge about realized output levels of firms' production and banks cannot observe these. To deal with such repeated informational friction we take the paper of Smith and Wang (2006) on "dynamic credit relationships in general equilibrium" as a starting point. Like them we have banks and ex-ante identical households with finite life expectancy who either become entrepreneurs or workers. Workers supply labor, consume and save, whereas entrepreneurs run firms by hiring labor and capital. The realized output of firms is exposed to stochastic states of productivity. These are only observable to the entrepreneurs who report them to the banks. We extend Smith and Wang (2006) by adopting a production structure which allows for variable firm size like in Clementi and Hopenhayn (2006) (who work in a partial-equilibrium analysis). In particular, we use a technology with decreasing returns to scale. As in Clementi and Hopenhayn (2006) and Smith and Wang (2006) entrepreneurs finance production costs through loans from banks. Banks offer entrepreneurs long-term financial contracts, which determine

the optimal level of bank loans and state-contingent repayments. In recursive formulation these are determined together with future promised values as functions of today's promised values. A promised value is the continuation utility of an entrepreneur from consumption of future cash flows (net revenue generated from production by using bank loans minus repayments). The financial contracts are promise keeping and incentive compatible and fulfill the limited liabilities and the credibility constraints.

Our model structure allows us to determine the share of entrepreneurs endogenously and to see the effects of the dynamic lending-contracts on the size, growth, variance of growth of firms at different ages, and on the size distribution of firms in the economy in equilibrium. This extends Smith and Wang (2006) by the aspect of firm dynamics and it completes Clementi and Hopenhayn (2006) by the general equilibrium aspect. Further, it adds to Dyrda (2016), Gross and Verani (2013) and Verani (2015) the endogenous determination of the share of entrepreneurs. They embed dynamic contract in a general equilibrium framework, but assume two types of households with different utility functions being either workers or entrepreneurs. We calibrate our model; use it to get numerical results and to solve for the general equilibrium. Workers' saving and labor decision, entrepreneurs' choice of the optimal level of factor inputs and the optimal financial contract are derived numerically. We find that the optimal level of bank loans and state-contingent future promised values are increasing functions of today's promised values while the state-contingent repayments first increase and then decrease with the state variable. State-contingencies of future promised values and repayments are as follows: If entrepreneurs report a high productivity state they are promised a higher future continuation utility, but they have to repay more today than if they report a low productivity state. This trade-off induces truth-telling about productivity realizations. By combining the three partial decision problems – of workers, entrepreneurs and banks, respectively – we close our model and determine the stationary general equilibrium. Our model predicts an equilibrium interest rate of around 4%. This is a common number in literature. The share of entrepreneurs in our economy is found to be 8%, which corresponds approximately to the rate of self-employed in the U.S. (data from OECD). The firm dynamics resulting in general equilibrium from the optimal path of the promised values are as follows: There is a positive correlation between firm size and firm age. Furthermore, the growth of younger firms is on average larger and more volatile than that of older firms.

In addition to the numerical results and the economic explanation of them, we provide a discussion of technical issues which can cause problems in dynamic programming. These are, among others, starting value problems, extrapolation issues, sensitivity to parameter values and to functional forms as well as issues related to simulations.

The paper adds to the literature by modeling the long-term credit relationships be-

tween firms and banks in general equilibrium. Through dynamic contracting information asymmetries between banks and firms can be overcome. The analysis of repeated information asymmetries was initiated by Radner (1985) and Rogerson (1985). The dynamic programming approach to it with the recursive formulation of incentive compatible, optimal contracts was developed by Green (1987) and Spear and Srivastava (1987). Thomas and Worrall (1990), who extend the two-period, two-state problem of Townsend (1982) for any number of periods and finite state spaces, add to Green (1987) and Spear and Srivastava (1987) by focusing on the long-run asymptotic properties of the contracts. Such incentive compatible long-term contracts deliver on the one hand an insurance component if agents are exposed to idiosyncratic shocks which are unobservable (as in Green (1987), Thomas and Worrall (1990), Atkeson and Lucas (1992) or Atkeson and Lucas (1995)). On the other hand, they provide financing opportunities. In particular, contracts between risk-neutral banks and firms can support optimal lending policies of banks which maximize the value of the firms (as in Quadrini (2004), Clementi and Hopenhayn (2006) or DeMarzo and Fishman (2007)).

The main contribution of this paper is the incorporation of dynamic financial contracts into a general equilibrium framework with an endogenous share of entrepreneurs and dynamic evolution of firms' size over age. Smith and Wang (2006), Dyrda (2016), Gross and Verani (2013) and Verani (2015) work in a similar model set-up. However, Smith and Wang (2006) do only consider projects with fixed units of capital and labor as input factors. They do not model an entrepreneur's optimal labor and capital decision under a more realistic production function. Hence, they cannot deal with firm dynamics. Dyrda (2016) does not consider a saving decision of workers by excluding them from the capital market. And in contrast to Dyrda (2016), Gross and Verani (2013) and Verani (2015), who do not incorporate firm entry, we determine the share of entrepreneurs in the economy endogenously. Technically, more complex equilibrium conditions are considered, which raises computational challenges.

Our model exhibits financial frictions which are the result of the information asymmetry. More precisely, we have borrowing constraints (i.e., firms do not get the efficient level of banks loans) as an endogenous result of the incentive-compatible long-term lending relationship between borrowers and the lender. This is like in the literature discussed above. Yet, in contrast to other contributions to the analysis of long-term contracts between firms and banks, our model does not connect financial frictions to the issue of collateral as it is done in Clementi and Hopenhayn (2006) or Verani (2015). Nor do we allow for the possibility of auditing like in Verani (2015) or Albuquerque and Hopenhayn (2004) in an environment with limited enforcement.

Directly connected to the (endogenous) borrowing constraints are the dynamics of

firm development in our model.[1] As a result of the long-term relation between the banks and firms, we predict that older firms are on average larger and that they grow less but more stable. These results are in line with the predictions from the dynamic contract models in Clementi and Hopenhayn (2006), Dyrda (2016), Gross and Verani (2013) and Verani (2015). Furthermore, they are consistent with the empirical regularities of firm dynamics.[2] In our model, the firm size distribution is more dispersed for older than for younger firms because their history of productivity realizations is more heterogeneous. That (endogenous) borrowing constraints have an impact on the size distribution of firms is consistent with the results of Angelini and Generale (2008) and Cabral and Mata (2003) who find that younger firms, which are financially constrained, have in fact different (more skewed) firm size distributions.

The structure of the paper is as follow: Section 1.2 introduces the theoretical model. Therein, the workers' and entrepreneurs' problems and the role of financial intermediaries is described. Further, the recursive formulation of the dynamic lending contracts is presented and some theoretical properties are discussed. Section 1.3 provides the aggregation and equilibrium conditions and defines the stationary, general equilibrium. In Section 1.4 the calibration of the model and numerical results are presented. In Section 1.5 we propose two model applications. Section 1.6 discusses issues connected with dynamic programming. Section 1.7 concludes.

1.2 Model

1.2.1 Model set-up

Consider an infinite time horizon model with finite life expectancy. A continuum of ex-ante identical households are born at the beginning of each period. A household survives at the end of the period with an exogenous probability. Right after birth a household decides to become a worker or an entrepreneur. We assume that this choice of occupation is irreversible over lifetime. A worker supplies labor, consumes and saves part of its income. An entrepreneur runs a firm which uses labor and capital as inputs and consumes entrepreneurial income (net revenue from production). In addition to the households, there are banks which act as financial intermediaries between workers and entrepreneurs. Namely, they take annuity deposits from workers and offer financing contracts in the form of bank loans to the entrepreneurs for their production. We assume that banks are

[1] For an overview of the effects of financial frictions in a dynamic contract set-up on aggregate fluctuations / business cycle fluctuations see the literature discussed in Dyrda (2016) and Verani (2015). For the effects of access to credit on international trade see Gross and Verani (2013).

[2] See, for example, Evans (1987) or Hall (1987) for empirical literature on firm dynamics.

competitive so that they make zero profit in expectation from any financial contract they sign with entrepreneurs.

1.2.2 Households

Households are endowed with one unit of labor each period and no wealth at birth. The instantaneous utility function of the households (both workers and entrepreneurs) is $U(c, l)$, where c is the consumption level and l is the labor supply. $U(c, l)$ is decreasing in l and increasing, strictly concave and bounded in c. Households discount future with rate β.

The exogenous survival probability is Δ. We assume that the mass of newborns in each period is $1 - \Delta$, so that the mass of population is constant at 1. The share of the households in cohort τ who become entrepreneurs is λ_τ. λ_τ will be determined endogenously in equilibrium by the labor market clearing condition (see 1.27). Figure 1.1 summarizes the compositions of different cohorts' population size and their occupations at time t. It shows that at each point in time we have a distribution of workers and of entrepreneurs of different ages in the population.

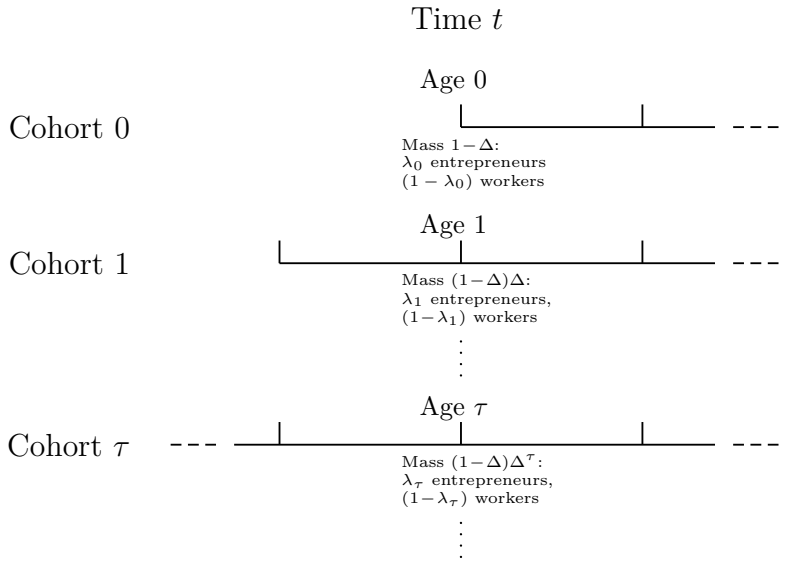

Figure 1.1: Size and occupational composition of different cohorts

1.2.2.1 Workers

In each period, workers supply labor for production and get wage income in return. Wage income as well as wealth can be used for consumption of final goods or as savings for wealth (and thus consumption) in future periods in the form of one-period annuity deposits in

the banks. In each period t, the workers of age τ buy at the end of the period $A_{\tau,t+1} \geq 0$ units of the annuity at price p_t^A. This entitles the worker to receive wealth level $A_{\tau,t+1}$ in period $t+1$ conditional on survival. The annuity deposits are priced competitively (actuarially fair) such that banks make zero profit from offering them to the workers. This means, the aggregate amount of money received by the banks from workers plus the interest it generates within a period must be equal to what they give out in the next period. Formally, at time t,

$$\sum_{\tau=0}^{\infty}(1+r_{t+1})(1-\Delta)\Delta^{\tau}p_t^A A_{\tau,t+1} = \sum_{\tau=0}^{\infty}(1-\Delta)\Delta^{\tau+1}A_{\tau,t+1},$$

where $(1-\Delta)\Delta^{\tau}p_t^A A_{\tau,t+1}$ is the aggregate payments of the workers of age τ at time t to buy the annuity. This generates an interest with rate $1+r_{t+1}$ in the next period. The aggregate amount is redistributed to all workers from last period who are still alive this period, which is Δ-times the original size $(1-\Delta)\Delta^{\tau}$ of each cohort. Therefore, zero-profit for banks implies that

$$p_t^A = \frac{\Delta}{1+r_{t+1}}, \tag{1.1}$$

whereby the market-clearing interest rate r_{t+1} is endogenously determined in equilibrium.

The workers' problem of choosing labor supply l, consumption c and savings in annuities A' in an optimal way can be formulated in the following recursive way with today' wealth $A \geq 0$ as state variable:

$$V^W(A;r,w) = \max_{c,l,A'}\left\{U(c,l) + \Delta\beta V^W(A';r',w')\right\}, \tag{1.2}$$

subject to

$$\begin{aligned} &c + p^A A' = wl + A, \\ &c \geq 0,\ l \in [0,1],\ A' \geq 0. \end{aligned} \tag{1.3}$$

$V^W(A;r,w)$ is the worker's value function (i.e., continuation utility) given today's wealth level A, interest rate r and wage rate w. p^A is given by (1.1). A prime indicates variables of tomorrow. $\Delta\beta$ captures discounting and the fact that the worker survives with probability Δ. Denote the policy function of optimal saving A' and labor choice l, respectively, by

$$A_{\tau,t+1} = g(A_{\tau,t};r_{t+1},w_t), \quad l_{\tau,t} = h(A_{\tau,t};r_{t+1},w_t). \tag{1.4}$$

1.2.2.2 Entrepreneurs

Entrepreneurs run firms. They supply entrepreneurial labor and derive utility from consumption of net revenue from production. Entrepreneurs and firms are associated for the whole lifetime. Namely, a newborn household who becomes entrepreneur opens a firm and runs the firm for the entire lifetime until death; then the firm exits the market. Thus, the firm's exit rate is exogenously given by the household's death rate $1 - \Delta$.

Firms produce in each period under uncertainty a single output (numéraire), which can either be consumed or be used as capital. In each period a fixed amount of entrepreneurial labor L^E is needed for setting up / managing the production. The production requires capital k and labor from workers l. The production function takes the form:

$$Y(k_t, l_t) = \theta_t F(k_t, l_t),$$

where $F(\cdot)$ reflects the production technology that transforms capital and labor inputs into the final product. It exhibits decreasing returns to scale. We assume the function to be continuous and strictly concave.

The level of θ_t represents the productivity at time t. In each period t the productivity is subject to an idiosyncratic shock with state space $\mathcal{S} = \{1, 2, \ldots, S\}$ and the corresponding realization of states $\theta_t \in \Theta = \{\theta_1, \theta_2, \ldots, \theta_S\}$. The shock is i.i.d. over entrepreneurs and time. The probability distribution of the states is $\{\pi_s\}_{s\in\mathcal{S}}$ with $\sum_{s\in\mathcal{S}} \pi_s = 1$. Without loss of generality, let $\theta_i < \theta_j$ if $i < j$. At any time t, each firm has an entire history of productivity realizations $\theta_\tau^t = (\theta_{t-\tau}, \ldots, \theta_{t-\tau+i}, \ldots, \theta_t)$, where τ is the age of the firm and $t - \tau + i$, $i \in \{0, 1, \ldots, \tau\}$ is the calendar time when the firm was of age i. Note that the heterogeneity among firms is characterized by the different histories of productivity realizations.

We assume that the realization of productivity shock is private information to the entrepreneur. This reflects the information asymmetry between entrepreneurs and banks.

Prior to production (i.e., before the idiosyncratic shock is realized), the entrepreneurs need to purchase capital and pay the workers. By assumption, the entrepreneurs are neither endowed with wealth nor do they accumulate wealth from their production revenues over lifetime. This means, self-financing of production is excluded. Hence, they need external financing. We restrict the source of financing to bank funding. Bank loans and repayments arise from a lifetime financial contract between the bank and the entrepreneur. More specifically, the financial contract entitles the entrepreneurs each period to some amount of bank loans b, which is used to cover the production costs, and some repayments m after production.[3] For a given level of loans and factor prices, the en-

[3] A detailed characterization of the financial contract, which includes bank loans b as well as repayments

trepreneurs determine the optimal capital and labor employment by maximizing expected output. The decision problem is

$$\max_{k_t,l_t} \mathbb{E}(\theta_t)F(k_t,l_t) \tag{1.5}$$

subject to

$$w_t l_t + (r_t+\delta)k_t \leq b_t,$$

where $(r_t+\delta)$ are the user cost of capital with δ being the depreciation rate of capital. We define

$$R(b_t;r_t,w_t) \equiv F(k_t^*,l_t^*) \tag{1.6}$$

with $k_t^* = k(b_t;r_t,w_t), l_t^* = l(b_t;r_t,w_t)$ being the solution to (1.5) at which the marginal rate of transformation correspond to the relative factor price of capital and labor. Notice that firms' labor costs include only wage payments to workers. The implicit assumption is that the entrepreneurs do not supply the entrepreneurial labor L^E in the labor market of workers. In what follows we denote the labor supply from workers as labor.

The entrepreneur's consumption c_t in each period is given by net revenue from production, which is gross production $\theta_t R(b_t)$ minus repayments to banks m_t:

$$c_t^E = \theta_t R(b_t;r_t,w_t) - m_t. \tag{1.7}$$

Therefore, the expected lifetime utility of an entrepreneur is given by

$$V_0^E = \sum_{t=0}^{\infty}(\beta\Delta)^t \mathbb{E}U(c_t^E, L^E), \tag{1.8}$$

where expectation is with respect to current period realization of productivity, θ_t, as well as the history of realizations captured in b_t and m_t (as derived in Section 1.2.4). According to the properties of the utility function, natural bounds for V_0^E are given by V_{min}^E and V_{max}^E, where

$$V_{min}^E \equiv \lim_{c\to 0}\frac{1}{1-\beta\Delta}U(c,L^E) \text{ and } V_{max}^E \equiv \lim_{c\to\infty}\frac{1}{1-\beta\Delta}U(c,L^E). \tag{1.9}$$

Remember that entrepreneurs do not make intertemporal savings decisions by assumption. Therefore, for given terms of the financial contract, maximization of expected lifetime utility in (1.8) is equivalent to maximizing the expected production output as given by (1.5).

In recursive formulation the continuation utility of an entrepreneur at time t can be

---

m, can be found in Section 1.2.4.

written as:

$$V_t^E = \mathbb{E}\left[U(c_t^E, L^E) + \beta \Delta V_{t+1}^E\right] \tag{1.10}$$

Notice that V_{min}^E and V_{max}^E are also the upper and the lower bound of continuation utilities of the entrepreneurs, respectively.

1.2.3 Financial intermediaries

The banks in the economy serve the role as financial intermediaries between saving households and producing firms. Namely, they take annuity deposits from workers and offer financial contracts to entrepreneurs. Banks act also as holder of capital. Banks' equity, E, is the accumulated retained earnings from net cash flows of bank loans and repayments (see Section 1.3.1.3 for more details on equity). Overall, banks invest the annuity deposits from workers and own banks' equity as capital input into the production run by entrepreneurs.

Banks are risk neutral profit maximizers and discount future at the current interest rate. There is free entry into the banking sector. This means in equilibrium banks expect zero profits from each single lending contract and thus size and ownership of the banks do not matter. Without loss of generality, we assume the existence of a representative bank holding a portfolio of all financial contracts with the entrepreneurs of all ages τ and with all heterogeneous histories of productivity realizations θ_τ^t.

1.2.4 Dynamic lending contract

The credit relation between banks and entrepreneurs is characterized by a lifetime binding financial contract. More specifically, following the standard dynamic contracting model (e.g., Thomas and Worrall (1990), Atkeson and Lucas (1992)), each firm signs a lifetime contract with a bank. Banks offer each newborn entrepreneur a take-it-or-leave-it lifetime binding financial contract.

We assume that both banks and entrepreneurs are fully committed to the contract in all possible future contingencies.

In the dynamic financial contract problem in recursive form, the continuation utility of an entrepreneur from future consumption, V_t^E as defined in (1.10), can be used as state variable (given interest and wage rate). Following the terminology of the literature, we call V_t^E the promised value. This means that the banks promise a continuation utility to the entrepreneurs by committing themselves to the terms of contract that imply a sequence of future consumption flows which generate the promised value. Therefore, given a promised value $V_t^E = V^E(\theta^{t-1}; r_t, w_t)$ as state variable – which includes the entire history of productivity realizations of an entrepreneur until time $t-1$, θ^{t-1} – the con-

tract consists of $\left\{b(V_t^E; r_t, w_t), m(V_t^E, \theta_t; r_t, w_t), V^E(V_t^E, \theta_t; r_t, w_t)\right\}$.[4] The first two terms are the bank loans to the entrepreneur, $b(V_t^E; r_t, w_t)$, and the repayments from the entrepreneur to the bank, $m(V_t^E, \theta_t; r_t, w_t)$. Loans are advanced before production, whereas repayments are made after the realization and after entrepreneurs' report of the current period productivity. Hence, loans are only contingent on today's promised value, whereas repayments are a function of today's promised value and the reported productivity level including time θ_t. (Note that according to the revelation principle, any equilibrium outcome can be achieved by a truth-telling mechanism. In particular, by imposing incentive constraints we can guarantee that entrepreneurs always report the actual realization of productivity θ_t. Therefore, we focus only on truth-telling contracts.) The third term, $V^E(V_t^E, \theta_t; r_t, w_t)$, is the next period's promised value given today's promised value and the productivity realization θ_t. In other words, $V_{t+1}^E = V^E(V_t^E, \theta_t; r_t, w_t)$ is the transition function of the state variable which incorporates the whole history of productivity realizations of an entrepreneur.

Since firms with the same promised value of today are assigned the same terms of contract (independent of time t or age τ), V^E is the state variable. V^E can be used as an indicator of firms in the equilibrium analysis (see Section 1.3.1.2 for aggregation of entrepreneurs).

1.2.4.1 Optimal financial contract

For given interest rate and wage rate, (r, w), the optimal contract can be determined by the following program written in recursive form with the promised value, $V^E \in [V_{min}^E, V_{max}^E]$, as state variable:

$$P(V^E; r, w) = \max_{b, \{m_s, V_s^E\}_{s\in\mathcal{S}}} -b + \sum_{s\in\mathcal{S}} \pi_s \left[m_s + \frac{\Delta}{1+r} P(V_s^E; r', w') \right] \tag{1.11}$$

subject to

$$V^E = \sum_{s\in\mathcal{S}} \pi_s [U\left(\theta_s R(b; r, w) - m_s, L^E\right) + \beta \Delta V_s^E], \tag{PK}$$

$$U\left(\theta_i R(b; r, w) - m_i, L^E\right) + \beta \Delta V_i^E \geq U\left(\theta_i R(b; r, w) - m_j, L^E\right) + \beta \Delta V_j^E, \ \forall i, j \in \mathcal{S}, \tag{IC}$$

$$m_s \leq \theta_s R(b; r, w), \forall s \in \mathcal{S}, \tag{LL}$$

$$V_s^E \in [V_{min}^E, V_{max}^E]. \tag{CC}$$

$P(V^E; r, w)$ is the bank's expected profit (value function) from a financial contract with state variable V^E given r and w. b denotes the level of bank loans, $\left\{m_s, V_s^E\right\}_{s\in\mathcal{S}}$ are state-

[4] See Appendix A.1 for a detailed structure of the timing in the dynamic financial contract.

contingent repayments and future promised values, respectively. $\frac{\Delta}{1+r}$ captures discounting and the fact that the entrepreneur survives with probability Δ. V^E_{min} and V^E_{max} are given in (1.9).

(PK) is the promise keeping constraint. It indicates that the terms of the contract must be such that the expected utility from today's cash flows plus future promised values fulfill the promised value V^E.

(IC) ensures that a contract is incentive compatible. Specifically, it guarantees that the truth-telling reporting strategy (weakly) dominates all other possible reporting strategies of entrepreneurs in terms of their expected utility, and thus eliminates incentives to misreport.

The constraints (LL) stand for limited liability. Since by assumption entrepreneurs do not own wealth, their liability for repayments to the bank are limited by the extent of the production revenue (i.e., realized productivity shock times the production level corresponding to the bank loan level). Hence, a contract is feasible if the terms of the contract are such that the entrepreneurs consume a non-negative amount of the final products after any productivity realization.

The credibility constraint (CC) imposes that banks could only promise utility values that are achievable with non-negative finite cash flows; otherwise, the promised value would only be granted by violating (LL) sometime in the future or is never satisfiable, respectively. More precisely, (CC) captures that banks can never promise (i) less utility than achievable by non-negative consumption for all future periods or (ii) more utility than by infinite consumption for all future periods.

Formally, we define an optimal financial contract as follows:

Definition 1.1. *For a given path* $\{r_t, w_t\}_{t=0}^{\infty}$*, the optimal dynamic contract is a sequence of functions* $\left\{b(V_t^E; r_t, w_t), m(V_t^E, \theta_t; r_t, w_t), V^E(V_t^E, \theta_t; r_t, w_t)\right\}_{t=0}^{\infty}$ *that solves program* (1.11)*.*

For notational simplicity we suppress from now r_t and w_t in the sequence of functions of the contract $\left\{b(V_t^E), m(V_t^E, \theta_t), V^E(V_t^E, \theta_t)\right\}_{t=0}^{\infty}$ and in the value function $P(V^E)$ whenever it is not misleading.

1.2.4.2 Theoretical properties

In this part, we show theoretical properties of financial contracts under program (1.11).[5] We first discuss general results about incentive compatible contracts and the simplification

[5]In the optimal dynamic contract, the path of factor prices r_t, w_t is taken as given. For notational simplicity we suppress from now on $\{r_t, w_t\}$ in the $R(b)$ function whenever it is not misleading.

of incentive constraints. Then, we come to the properties of the optimal contract. Proposition 1.1 and 1.2 and Lemma 1.1 and 1.2 follow the properties of optimal social insurance in Ljungqvist and Sargent (2000) which are based on Thomas and Worrall (1990).

The following proposition defines the necessary condition of an incentive compatible contract:

Proposition 1.1. *Let $\theta_s > \theta_{s-1}, \forall s \in \mathcal{S}$. An incentive compatible contract satisfies $m_s \geq m_{s-1}$ and $V_s^E \geq V_{s-1}^E$.*

Proof. See Appendix A.2.1.1. □

This implies that banks induce truth-telling behavior of entrepreneurs by postponing rewards for reporting high productivity realization. If productivity is high repayments are high, but the future promised value is high, too.

Define the incentive constraints for all $i, j \in \mathcal{S}$ as:

$$C_{i,j} \equiv U(\theta_i R(b) - m_i, L^E) + \beta \Delta V_i^E - U(\theta_i R(b) - m_j, L^E) - \beta \Delta V_j^E \geq 0, \tag{1.12}$$

where i is the actual state and j is the reported state. Then, the set of incentive constraints can be simplified with the following lemma.

Lemma 1.1. *If the local downward constraints, $C_{s,s-1} \geq 0$, and the local upward constraints, $C_{s,s+1} \geq 0$, hold for each $s \in \mathcal{S}$, then the constraints $C_{i,j} \geq 0$ hold $\forall i, j \in \mathcal{S}$.*

Proof. See Appendix A.2.1.2. □

Suppose for the following lemma and Proposition 1.2 and 1.3 that $P(V^E)$ is strictly concave – a fact which is observed in the numerics.

Using this and Lemma 1.1, we get the following property of the optimal contract.

Lemma 1.2. *For strictly concave $P(V^E)$, for all states $s \in \mathcal{S}$, the optimal contract implies that the local downward constraints $C_{s,s-1} \geq 0$ always bind, whereas the local upward constraints $C_{s-1,s} \geq 0$ never bind for $m_s > m_{s-1}$.*

Proof. See Appendix A.2.1.3. □

In addition, the optimal contract has the property of risk sharing:

Proposition 1.2. *For strictly concave $P(V^E)$, both the entrepreneurs' utility and the banks' profits are non-decreasing with a higher productivity realization, that is: Under an optimal contract, for $\theta_i > \theta_j$*

$$U(\theta_i R(b) - m_i, L^E) + \beta \Delta V_i^E \geq U(\theta_j R(b) - m_j, L^E) + \beta \Delta V_j^E, \tag{1.13}$$

$$-b + m_i + \frac{\Delta}{1+r} P(V_i^E) \geq -b + m_j + \frac{\Delta}{1+r} P(V_j^E). \tag{1.14}$$

Proof. See Appendix A.2.1.4. □

Next, we introduce the efficient level of bank loan, b^*, which is implicitly determined by

$$\mathbb{E}(\theta)R'(b^*; w, r) = 1, \tag{1.15}$$

that is, marginal productivity equals marginal costs of one more unit of bank loans. Notice that the efficient level of bank loans corresponds to the optimal firm size if banks were the firm owners.

Suppose for Proposition 1.3 that there are only two states in the state space, $\mathcal{S} = \{l, h\}$ with $\theta_h > \theta_l$.

Proposition 1.3. *For strictly concave $P(V^E)$ and for $m_s > m_{s-1}$, the optimal level of bank loans from the contract is not larger than the efficient level.*

Proof. See Appendix A.2.1.5. □

Note that this implies endogenous borrowing constraints, which are also existent in the models of Clementi and Hopenhayn (2006), Dyrda (2016), Gross and Verani (2013) and Verani (2015).

1.3 Aggregation and general equilibrium

So far we have characterized the optimization problems of the agents in the economy. More specifically, for a given sequence of factor prices $\{r_t, w_t\}_{t=0}^{\infty}$ and the share of entrepreneurs of each cohort $\{\lambda_\tau\}_{\tau=0}^{\infty}$, we get: (i) the workers' optimal path of consumption, wealth accumulation and labor supply from (1.2); (ii) the entrepreneurs' optimal path of capital and labor employment from (1.5); and (iii) the banks' optimal path of terms of contract with loans, repayments and future promised values from (1.11). Given the technical complexities, for combining the three partial parts to get the general equilibrium we focus on the stationary case with constant factor prices $\{r, w\}$ and a constant share of entrepreneurs λ. Specifically, we consider the age-dependent, time-independent supplies and demands of labor and capital, consumption of workers and entrepreneurs, bank loans and repayments. We sum the individual decisions over the cohorts of all ages in the economy to get the aggregate demand and supply of labor, capital and goods. This allows us in the end to write down the equilibrium conditions and define the general equilibrium. More precisely, the equilibrium is then the prices $\{r, w\}$ and the share of entrepreneurs λ such that goods, labor and capital markets clear and banks make zero profit (see Sections 1.3.2 and 1.3.3).

1.3.1 Aggregation

1.3.1.1 Aggregation of workers

By aggregating the optimal consumption, saving and labor decision over individual workers of all ages τ, we get total consumption C^W, total deposits D and total labor supply L^S:[6]

$$C^W(r,w) = \sum_{\tau=0}^{\infty}(1-\Delta)\Delta^\tau c(A_\tau, r, w) \tag{1.16}$$

$$D(r,w) = \sum_{\tau=0}^{\infty}(1-\Delta)\Delta^\tau p^A A_{\tau+1} = \sum_{\tau=0}^{\infty}(1-\Delta)\Delta^\tau p^A g(A_\tau, r, w) \tag{1.17}$$

$$L^S(r,w) = \sum_{\tau=0}^{\infty}(1-\Delta)\Delta^\tau l_\tau = \sum_{\tau=0}^{\infty}(1-\Delta)\Delta^\tau h(A_\tau, r, w) \tag{1.18}$$

where $c(\cdot)$ is given by the workers' budget constraint (1.3) and $g(\cdot)$ and $h(\cdot)$ are the worker's policy functions defined in (1.4). $(1-\Delta)\Delta^\tau$ is the mass of households of age τ. Note that heterogeneity among workers comes only from age differences; within a cohort all workers are identical in their lifetime decisions.

1.3.1.2 Aggregation of entrepreneurs

Aggregating over all entrepreneurs is more complicated because they are heterogeneous in two dimensions: Age and history of productivity realizations. In other words, there are firms of different ages τ and firms of the same age τ differ in productivity history θ^τ due to the idiosyncratic shocks.

History of productivity realizations of an entrepreneur aged τ, $\theta^\tau \in \Theta^\tau$ maps into a promised value V^E by applying the transition function $V_s^E = V^E(V^E, \theta_s; r, w)$ recursively with starting value V_0^E.[7] The distribution of θ^τ among entrepreneurs of age τ corresponds to a stationary distribution of promised values, denoted by $\Psi_\tau(V^E)$.

Promised values V^E are translated by the optimal financial contract into bank loans and repayments, $\left\{b\left(V^E; r, w\right), m\left(V^E, \theta_s; r, w\right)\right\}$. For given bank loans, the optimal capital and labor employment, $\left\{k^*(V^E; r, w), l^*(V^E; r, w)\right\}$ are given by the solution to (1.5) where it is used that $b(V^E; \cdot)$ is a function of V^E.

We can aggregate bank loans B, capital K^D and labor demand L^D over all cohorts as

[6] For now the measure of workers is supposed to be 1. The equilibrium share of workers $(1-\lambda)$ will be determined through the equilibrium conditions as given in Section 1.3.2.

[7] The indifferent occupational choice condition, which must hold in equilibrium, requires that $V_0^E = V^W(0; r, w)$ (see (1.26)).

follows:[8]

$$B(r,w) = \sum_{\tau=0}^{\infty}(1-\Delta)\Delta^{\tau}\int b(V^E;r,w)d\Psi_{\tau}(V^E), \tag{1.19}$$

$$K^D(r,w) = \sum_{\tau=0}^{\infty}(1-\Delta)\Delta^{\tau}\int k^*(V^E;r,w)d\Psi_{\tau}(V^E) \tag{1.20}$$

$$L^D(r,w) = \sum_{\tau=0}^{\infty}(1-\Delta)\Delta^{\tau}\int l^*(V^E;r,w)d\Psi_{\tau}(V^E) \tag{1.21}$$

Furthermore, the aggregate expected repayments from the entrepreneurs of all ages τ to banks are given by:

$$M(r,w) = \sum_{\tau=0}^{\infty}(1-\Delta)\Delta^{\tau}\sum_{s\in S}\pi_s\int m(V^E,\theta_s;r,w)d\Psi_{\tau}(V^E) \tag{1.22}$$

In a similar way, the expected aggregate output Y and the consumption of the entrepreneurs C^E are given by:

$$Y(r,w) = \sum_{\tau=0}^{\infty}(1-\Delta)\Delta^{\tau}\sum_{s\in S}\pi_s\int \theta_s R(b(V^E;r,w);r,w)d\Psi_{\tau}(V^E), \tag{1.23}$$

$$C^E(r,w) = \sum_{\tau=0}^{\infty}(1-\Delta)\Delta^{\tau}\sum_{s\in S}\pi_s\int c(V^E,\theta_s;r,w)d\Psi_{\tau}(V^E), \tag{1.24}$$

where $R(\cdot)$ is defined in (1.6) and $c(\cdot)$ in (1.7).

1.3.1.3 Aggregation of banks' equity

Finally, banks' equity is the accumulated retained earnings from the flows of bank loans and repayments. In a stationary equilibrium, it is determined by

$$E(r,w) = (1+r)E(r,w) + M(r,w) - B(r,w),$$

where $E(r,w)$ denotes the bank equity, $(1+r)E(r,w)$ are the gross returns on previous equity and $M(r,w) - B(r,w)$ are the net aggregate payments from a measure 1 of

[8]For now the measure of entrepreneurs is supposed to be 1. The equilibrium share of entrepreneurs λ will be determined through the equilibrium conditions as given in Section 1.3.2.

entrepreneurs.[9] Rewriting the above equation, we have

$$E(r,w) = \frac{B(r,w) - M(r,w)}{r}. \tag{1.25}$$

1.3.2 Equilibrium conditions

In a stationary equilibrium (see Section 1.3.3 for the formal definition), there are simultaneously workers and entrepreneurs from all cohorts in the economy. Newborn households are indifferent with respect to their occupational choice. That is, the expected lifetime utility of becoming a worker is the same as that of becoming an entrepreneur. Formally, this means:

$$V_0^E = V^W(0;r,w), \tag{1.26}$$

where $V^W(0;r,w)$ is determined by program (1.2). Note that this equation defines the starting value of the state variable for the entrepreneurs, which is used in generating the life path of promised values in the numerical analysis by applying the optimal contracts (see Section 1.4.2).

In addition – as in standard general equilibrium theory – labor, capital and goods markets clear.

Labor market clearing requires that aggregate labor supply from workers equals aggregate demand for labor by the entrepreneurs. This is

$$\lambda L^D(r,w) = (1-\lambda)L^S(r,w), \tag{1.27}$$

with $L^S(r,w)$ and $L^D(r,w)$ defined in (1.18) and (1.21), respectively, and λ being the endogenously determined share of the entrepreneurs in the economy.

Capital market clearing requires in equilibrium that capital supply in the economy, which consists of aggregate deposits from the workers plus banks' equity, is equal to capital demand:

$$K^S(r,w) \equiv (1-\lambda)D(r,w) + \lambda E(r,w) = \lambda K^D(r,w), \tag{1.28}$$

where $D(r,w)$, $E(r,w)$ and $K^D(r,w)$ are given in (1.17), (1.25) and (1.20), respectively.

The goods market is cleared if aggregate output equals the sum of households' consumption plus aggregate investments, where the latter is equal to depreciated capital in

[9] Notice that this condition indicates that in the stationary equilibrium banks give on aggregate more loans than repayments they ask for; with the gap between B and M being exactly coverable by the interest from banks' equity. Thus, the level of equity is endogenously kept constant in the stationary case. Since we do not characterize the path of how the economy converges to the stationary equilibrium, we cannot show numerically how the accumulation of banks' equity converges to the stationary equilibrium level. However, we give in Appendix A.5 a non-rigorous intuition of how an economy may evolve from the very beginning of time to the stationary equilibrium.

a stationary equilibrium. Formally, the condition is

$$\lambda Y(r,w) = (1-\lambda)C^W(r,w) + \lambda C^E(r,w) + \delta\lambda K^D(r,w). \tag{1.29}$$

It is directly implied by the labor and the capital market clearing conditions, (1.27) and (1.28) as shown Appendix A.2.2.

Finally, banks' are assumed to make zero profit in expectation from each newly-signed contract in equilibrium. Under the indifferent occupational choice condition in (1.26), the zero-profit condition for banks is given by

$$P(V^W(0;r,w)) = 0. \tag{1.30}$$

1.3.3 Definition of general equilibrium

With the agents' optimal behavior derived from the respective optimization problems and the general equilibrium conditions, we can now define the stationary general equilibrium in the economy.

Definition 1.2. *A stationary general equilibrium is characterized by a stationary distribution of workers of different ages, and the corresponding capital and labor supply* $\{A_\tau, l_\tau\}_{\tau=0}^{\infty}$*, a stationary distribution of entrepreneurs of different ages, for each cohort a stationary distribution of promised values,* $\left\{\Psi_\tau(V^E)\right\}_{\tau=0}^{\infty}$*, and the corresponding capital and labor demand of the entrepreneurs,* $\left\{k^*(V^E), l^*(V^E)\right\}$*, bank loans and repayments of the banks,* $\left\{b(V^E), m(V^E,\theta_s)\right\}_{s\in\mathcal{S}}$*, and interest rate, wage rates and share of entrepreneurs,* $\{r,w,\lambda\}$ *such that for given* (r,w)*,*

(1) workers maximize lifetime utility according to (1.2)*,*

(2) entrepreneurs maximize expected output according to (1.5)*,*

(3) banks offer profit-maximizing contracts subject to (PK), (IC), (LL), (CC) according to (1.11)*.*

The factor prices (r,w) *and share of entrepreneurs* λ *are such that,*

(1) labor, capital and goods market clear according to (1.27)*,* (1.28) *and* (1.29)*.*

(2) banks make zero profit in expectation according to (1.30)*.*

We determine the stationary equilibrium numerically, but we do not deliver an analytical general proof for the existence of a stationary equilibrium.[10]

[10]See Appendix A.3.3for a detailed description of the algorithm to find the stationary equilibrium numerically.

1.4 Calibration and numerical results

Given the complexity of the problem, the stationary equilibrium is determined numerically in the following. To calibrate the model, we assume specific functional forms of the utility and the production function and give exogenous parameter values.

The households' utility function (workers and entrepreneurs) is given by

$$U(c,l) = -exp(-\gamma c) - \eta l^2, \quad \gamma, \eta > 0. \tag{1.31}$$

It includes a CARA-part for consumption with γ being the absolute risk aversion and a parabola part for the disutility of labor supply. The form of the utility function gives us computational simplicity.

The production technology of the entrepreneurs exhibits decreasing return to scale:

$$Y(k,l) = \theta_s \bar{a} k^{\alpha_k} l^{\alpha_l}, \tag{1.32}$$

where θ_s denotes the state-dependent productivity realization, $\bar{a}$ scales total factor productivity and α_k and α_l are the share of capital and labor, respectively. We simplify the state space $\mathcal{S}$ to two states: "high" and "low" with productivity $\theta_h = \theta + \sigma$ and $\theta_l = \theta - \sigma$, $\sigma > 0$, and corresponding probability π_h and π_l, respectively.

For the exogenous parameters we take the values given in Table 1.1. The survival rate

Table 1.1: Exogenous parameters

Parameters		Value
Survival rate	Δ	0.92
Discount rate	β	0.963
Household preferences	γ	2
	η	0.5
Probability of bad state	π_l	0.5
High productivity	θ_h	1.25
Low productivity	θ_l	0.75
Fixed entrepreneur labor	L^E	1/3
Share of capital	α_k	0.35
Share of labor	α_l	0.6
Productivity scale	$\bar{a}$	1/3
Depreciation rate	δ	0.1

is chosen such that the death rate $1-\Delta$ corresponds approximately to the empirical yearly exit rate of firms. The discount rate β is similar to standard values found in literature. Household preference parameter γ and η are internally calibrated such that workers'

labor supply is about 30% of their labor endowment. Further, we assume both states are equally likely. Then, the values of θ_h and θ_l imply an expected productivity realization of $\theta = 1$, with standard deviation of 0.25. L^E corresponds to a third of an entrepreneur's labor endowment. α_k and α_l correspond to the capital and labor shares of output. The depreciation rate $\delta = 0.1$ corresponds to a common number in literature reflecting a quarterly depreciation rate of approximately 2.5%. The assumed utility function and the parameter values determine the boundaries of the promised value, $V^E_{min} = -9.26$ and $V^E_{max} = -0.49$ given by (1.9).

1.4.1 Three optimization problems

We characterize first the numerical solutions to the three optimization problems. Specifically, for a given wage w and interest rate r, we solve for the workers' optimal consumption, saving and labor supply decision based on (1.2), the entrepreneurs' capital and labor demand as in (1.5) and especially the banks' optimal financial contract from (1.11).[11]

1.4.1.1 Workers' optimal decisions

Figure 1.2 depicts, as a function of the current period deposit wealth A, the workers' optimal consumption $c(A)$, the labor supply $l(A)$ and the saving decision $A'(A)$ corresponding to the policy functions given in (1.4) and the lifetime expected utility $V^W(A)$ for given w and r.[12]

They are in line with the results from standard lifetime utility maximization: Households consume more today and save more for tomorrow if their current wealth A is higher. One has $A'(A) > 0$ for all A, which means that households always decide to hold positive annuity deposits. Further, with more A they supply less labor because they are less dependent on labor income. Their lifetime expected utility, captured by the value function $V^W(A)$, is an increasing function in A, indicating that workers are better off if endowed with more wealth A.

[11] In the figures, we use $w = 0.1599$ and $r = 0.0417$, which are the equilibrium values later determined numerically in the general equilibrium in Section 1.4.2 by using the search algorithm described in Appendix A.3.4. For simplicity we suppress w and r in the notation.

[12] See Appendix A.3.1 for the procedure to solve the recursive workers' problem numerically.

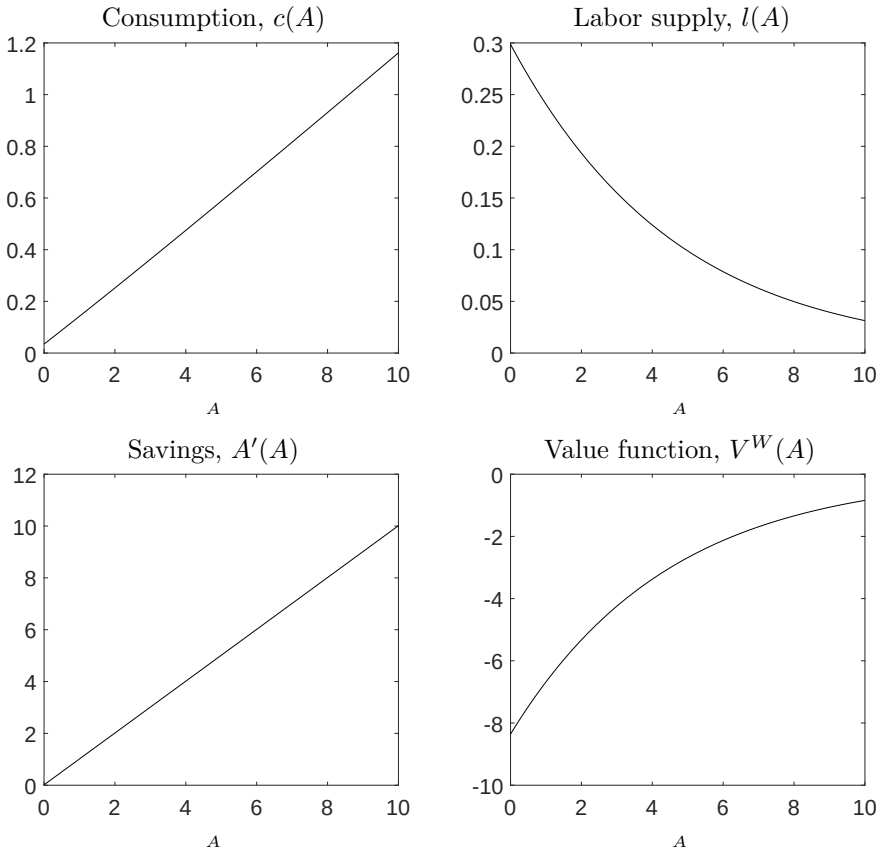

Figure 1.2: Solution to worker's problem

1.4.1.2 Entrepreneurs' optimal capital and labor employment

For a given level of bank loans b and factor prices w and r, the entrepreneur chooses optimally capital input and labor employment based on the decision problem in (1.5) as follows:

$$k^* = \frac{1}{r+\delta}\frac{\alpha_k}{\alpha_k+\alpha_l}b \quad \text{and} \quad l^* = \frac{1}{w}\frac{\alpha_l}{\alpha_k+\alpha_l}b. \tag{1.33}$$

Hence,

$$R(b) = \left(\frac{\alpha_k}{r+\delta}\right)^{\alpha_k}\left(\frac{\alpha_l}{w}\right)^{\alpha_l}\left(\frac{b}{\alpha_k+\alpha_l}\right)^{\alpha_k+\alpha_l}\bar{a}. \tag{1.34}$$

Following from equation (1.15), the efficient level of bank loans is thus given by

$$b^* = (\alpha_k+\alpha_l)\left[\bar{a}\left(\frac{\alpha_k}{r+\delta}\right)^{\alpha_k}\left(\frac{\alpha_l}{w}\right)^{\alpha_l}\right]^{\frac{1}{1-\alpha_k-\alpha_l}}. \tag{1.35}$$

Figure 1.3 shows this capital and labor demand of entrepreneurs as function of the bank loans b for given r and w. Capital and labor demand are linearly increasing functions in b. For the given form of the production function, the capital intensity is independent

of the level of the bank loan b.

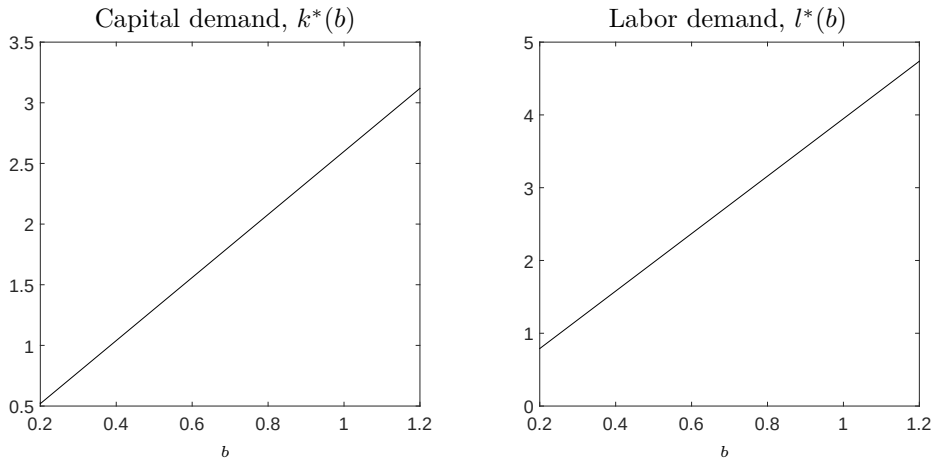

Figure 1.3: Solution to entrepreneur's problem

The outcomes indicate that the more bank loans firms get, the more input they are demanding. This implies that the size of production increases in the amount of available funds in the form of bank loans. Hence, bank loans determine the size of firms. Following this we will later use the amount of bank loans as the indicator of firm size and discuss based on this the dynamics of average firm size, growth and variance of growth at different ages of the firms (see Section 1.4.4).

1.4.1.3 Banks' optimal financial contract

Figure 1.4 shows (for given r and w) as a function of today's promised value V^E (state variable), the banks' profit $P(V^E)$, state-contingent future promised value $V_s^E(V^E)$, state-contingent repayments $m_s(V^E)$ and the bank loans $b(V^E)$.[13] State-contingency is captured by the subindex, $s \in \{l, h\}$, with l and h standing for low and high productivity realizations, respectively.

The banks' profit is $P(V^E)$ is strictly concave. For V^E not close to V_{min}^E, $P(V^E)$ is clearly decreasing in V^E.

The state-contingent future promised values, $V_l^E(V^E)$ and $V_h^E(V^E)$ are strictly increasing in V^E. Further, one can see from the subplot of $V_s^E(V^E)$ that $V_l^E < V^E$ and $V^E < V_h^E$. For values of V^E very close to V_{min}^E the lower credibility constraint (CC) is binding. In other words, without imposing the credibility constraint (CC), $V_l^E(V^E) < V_{min}^E$ would result for values of V^E very close to V_{min}^E, which contradicts $c \geq 0$ sometime in future.[14]

[13] See Appendix A.3.2 for the numerical procedure to solve the recursive formulated lending contract.

[14] This shows that accounting the credibility constraints is essential.

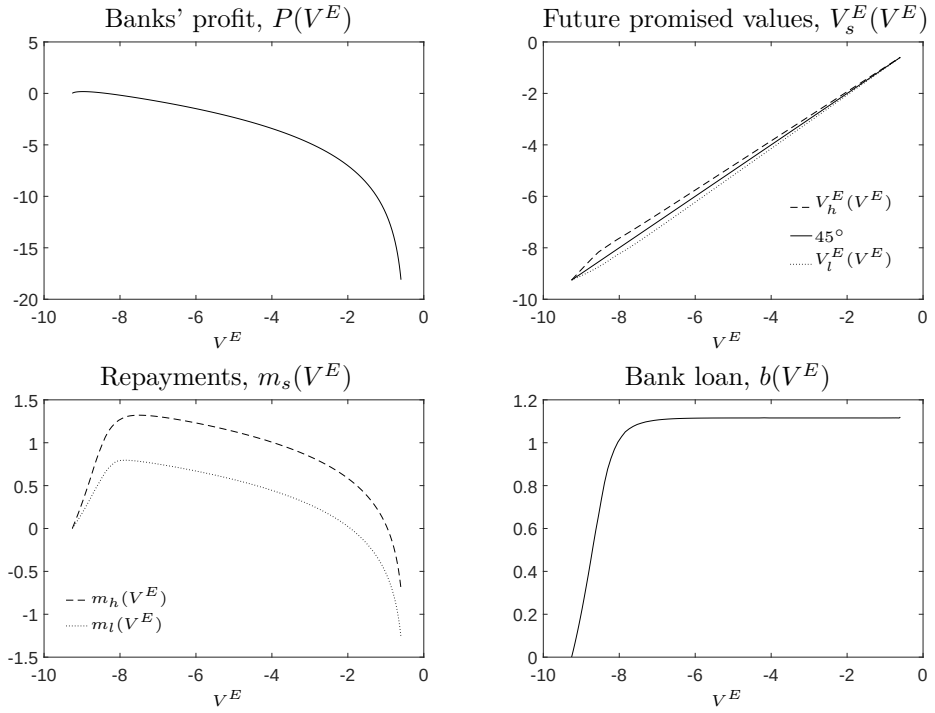

Figure 1.4: Optimal contract

State-contingent repayments $m_s(V^E)$, $s \in \{h, l\}$ are non-monotonic; repayments $m_s(V^E)$ first increase in V^E and then decreases at higher promised values.[15] The latter means, firms with a high promised value V^E have to repay less (even $m_s(V^E) < 0$) with the intuition that otherwise high V^E could not be realized without exploding V^E-path. Further, $m_l < m_h$ says that firms with a low productivity shock are spared from high repayments.

Note that the two subplots $V_s^E(V^E)$ and $m_s(V^E)$ for $s \in \{h, l\}$ reflect the theoretical results (see Proposition 1.1): Importantly, a postponed reward for reporting a high productivity state (high m_h, high V_h^E) and a postponed punishment for reporting a low productivity state (low m_l, low V_l^E) provide the entrepreneurs incentive to report the actual productivity realization.

Figure 1.4 shows further that the level of bank loan $b(V^E)$ is strictly increasing in V^E.[16] By comparing the level of bank loans $b(V^E)$ with the expected repayment $\pi_l m_l(V^E) + (1 -$

[15]Non-monotonicity can arise as a result of the functional forms of the utility, the production and the profit function, and their relative curvature compared to each other; the banks fulfill higher promised values V^E by both higher future promised utility and higher current consumption (through b to m_s).

[16]The specific shape of $b(V^E)$ is the result of the functional forms of the utility and the production function and their relative curvature compared to each other (see (A.9) in Appendix A.2.1). There are unstable $b(V^E)$ for V^E-values approaching V_{max}^E due to computational difficulties for values close to V_{max}^E. However, for determining the equilibrium this problem is negligible because firms hardly reach promised V^E-values in the region close to V_{max}^E when starting at $V_0^E = -8.36$ as derived in the general equilibrium (e.g., 65 years of always high productivity shock, which would leads to $V^E > -1$ has probability $(\Delta(1 - \pi_l))^{65} = 1.2 \cdot 10^{-22} \approx 0$). Further, the highest V^E reached by an entrepreneur in the simulation of our economy is only -1.51.

$\pi_l)m_h(V^E)$ (see Appendix A.5), one sees that for low V^E the expected repayments exceed the level of bank loans. Thus, banks retain earnings from the contracts at such state variable levels. For higher V^E the reverse holds which means that entrepreneurs retain deposited resources. For all V^E, $b(V^E)$ is smaller than the efficient level $b^* = 1.1814$ defined in (1.15); in line with Proposition 1.3. The increasing function $b(V^E)$ means that firms with a higher promised value V^E get more bank loans and are thus larger. Hence, the transition function of the promised value, $V_s^E(V^E)$, is crucial in generating firm dynamics: For given current period productivity realization, the future promised value to entrepreneurs, $V_s^E(V^E)$, determines the level of tomorrow's bank loans and thus the evolution of the firm size. The relative level of bank loans available to a firm in two successive periods given by $\frac{b(V_s^E)}{b(V^E)}$, $s \in \{h, l\}$ depends on the productivity realization: A high productivity shock entitles the firm to more bank loans in the next period while a low productivity shock lowers b (see Figure 1.5, which gives a similar pattern as in Clementi and Hopenhayn (2006)).[17]

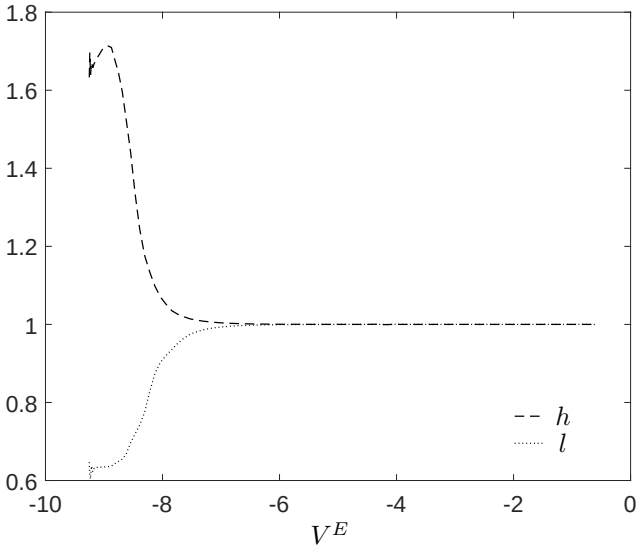

Figure 1.5: Relative change of bank loans, $\frac{b(V_s^E)}{b(V^E)}$, $s \in \{h, l\}$

[17]The pattern of the two curves is determined by two factors: The gap between today's and tomorrow's promised value, V^E and $V^s(V^E), s \in \{h, l\}$, and the level of bank loans, $b(V^E)$. At low level of promised value today (i.e., small V^E), the gap between V^E and $V^s(V^E)$ is relatively large. In addition, the level of bank loans is very sensitive to change in promised values. The combination of the two leads to a large relative change of bank loans between today and tomorrow, $\frac{b(V_s^E)}{b(V^E)}$. At high level of V^E, however, level of bank loans is almost flat; change in promised values has nearly no impact on the level of bank loans a firm gets. In addition, tomorrow's promised values (in both states) are very close to that of today. Therefore, the relative change of bank loans in both states approaches 1.

1.4.2 General equilibrium

With the solutions of the three optimization problems, we can now determine the general equilibrium in our economy. The stationary equilibrium values of the endogenous factor prices (r, w) and the share of entrepreneurs λ are simultaneously found by labor and capital market clearing and banks' zero-profit condition. Thus, for determining the equilibrium, aggregate demands and aggregate supplies of labor and capital and the starting promised value $V^W(0; r, w)$ must be calculated.

The labor supply and part of the capital supply come from workers. In the stationary equilibrium, aggregating total deposits D and total labor supply L^S of all generations in the economy is computationally equivalent to the aggregation of deposits and labor supply of one cohort over its lifetime as is implied by (1.17) and (1.18). The weights $(1-\Delta)\Delta^\tau$ correspond then to the size of the cohort at the different ages τ. Since workers are homogeneous within cohorts with identical saving and working decision, it is numerically straightforward to compute the aggregate savings and labor supply using (1.17) and (1.18).[18]

To derive the demand for labor and capital we simulate life paths of entrepreneurs with stochastic shocks in their productivity and exogenous death.[19] Thus, we have simulated the length of life for each entrepreneur and its history of productivity realizations, $\theta^t = \{\theta_1, \theta_2, \dots \theta_t\}$. Starting at promised value $V_0^E = V^W(0, r, w)$, the simulated history of productivity realizations generates then for each entrepreneur a lifetime sequence of promised values $\left\{V_1^E, V_2^E, \dots, V_t^E\right\}$ by applying the transition function $V_s^E(V^E)$ recursively.[20] To the sequence of $\left\{V_i^E\right\}_{i=1}^t$ correspond directly a sequence of repayments $\{m_i\}_{i=1}^t$ and a sequence of bank loans $\{b_i\}_{i=1}^t$. Aggregating these at t over all entrepreneurs we get in the end total repayments M and total bank loans B. The latter determines total labor and capital demands L^D and K^D: Because of the linear relation between b and k^*, b and l^* according to (1.33), the total labor and capital demands defined in (1.20) and (1.21), are given respectively by

$$K^D = B\frac{1}{r+\delta}\frac{\alpha_k}{\alpha_k+\alpha_l}, \text{ and } L^D = B\frac{1}{w}\frac{\alpha_l}{\alpha_k+\alpha_l}. \tag{1.36}$$

[18] See step 4 in Appendix A.3.3 for the aggregation of the supply side.

[19] See step 1 in Appendix A.3.3 for the description of the simulation procedure with $N^E = 10,000,000$ life paths.

[20] See Figure A.4-A.6 in Appendix A.6 for three different examples of life paths of a 50 year old entrepreneur: Life path I illustrates a lucky life with many high productivity shocks. Life path II represents a life with a relatively balanced history of productivity realizations and life path III was driven by bad luck with many low productivity shocks. One can see that high productivity shocks tend to increase V^E overtime, while low productivity shocks lower it. The transition of V^E translates directly into the evolution of b and m.

To determine the general equilibrium, we now use these aggregate demands and supplies. The share of entrepreneurs λ is determined by the labor market clearing condition (1.27). Namely, $\lambda = \frac{L^S}{L^D+L^S}$. The factor prices w and r are simultaneously determined by the capital market clearing condition (1.28) and the expected zero profit condition (1.30).[21] The resulting equilibrium values of r, w and λ are shown in Table 1.2.

Table 1.2: Equilibrium parameter

Parameters		Value
Interest rate	r^*	4.17%
Wage	w^*	0.1599
Share of entrepreneurs	λ^*	7.62%

Our equilibrium interest rate is around 4%, which is a common number in literature. The equilibrium share of entrepreneurs λ is 7.6% and corresponds approximately to the rate of self-employed labor of around 7% in the U.S. over the last years (data from OECD).

The equilibrium lifetime expected utility and other equilibrium values are given in Table 1.3.

Table 1.3: Equilibrium values

Parameters		Value
Lifetime utility	$V_0^E = V^W(0, r^*, w^*)$	-8.3549
Total labor supply	L^S	0.2879
Total capital supply	D	0.1618
Total bank loans	B	0.8836
Total labor demand	L^D	3.4906
Total capital demand	K	2.2972
Total repayments	M	0.8695

The lifetime utility of entrepreneurs and workers is $V_0^E = V^W(0, r^*, w^*) = -8.36$. The total labor supply L^S corresponds to about a third of a worker's labor endowment, which is in line with standard values from the empirics. Further, from the amount of bank loans B and repayments M given in Table 1.3 we can calculate the amount of banks' equity E using (1.25). This indicates is an equity ratio $E/K = 14.68\%$. This number is above current levels of large international banks, but below the proposed level of 20% by Admati and Hellwig (2013).

[21] See step 5 in Appendix A.3.3 for the procedure to determine the equilibrium in which the two conditions are jointly fulfilled, and Appendix A.3.4 and A.4 for the detailed description of the algorithm to find the stationary equilibrium numerically. We approximate the labor market clearing up to a residual of magnitude 0, the residual in the capital market is -0.00015 and the deviation from the zero-profit condition is -0.00016.

1.4.3 Firm distributions

In this equilibrium, we can derive distributions for firm characteristics from the simulation of the paths of the entrepreneurs' lives. Figure 1.6 shows the distribution of entrepreneurs in the economy with respect to different characteristics: Age, promised values, repayments and bank loans.

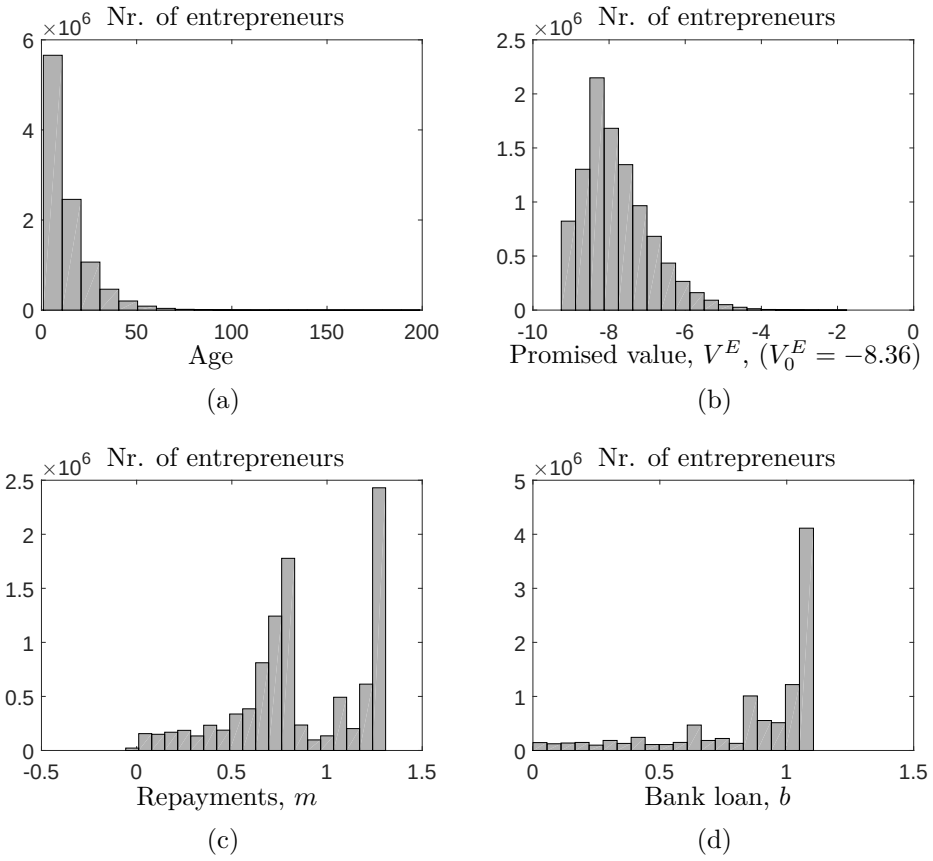

Figure 1.6: Distribution of age, promised values, repayments and bank loans

Subplot (a) shows the distribution of entrepreneurs' ages. With a share of $1-\Delta = 8\%$ most entrepreneurs are newborns. Then, one-year old represent a share of $(1-\Delta)\Delta = 7.36\%$ and so on. Finally, the share of entrepreneurs older than 50 years account for only 0.16% in our economy.

Subplot (b) shows the distribution of promised values V^E. We get the histogram of the distribution of firm promised values $\Psi(V^E)$ as shown in Subplot (b) by counting the number of entrepreneurs in the economy in different bins of $V^E \in [V_{min}, V_{max}]$. The plot indicates clearly that the mass of the promised values lies around the starting value $V_0^E = -8.36$. Firm heterogeneity then arises from the different length and composition of productivity realizations over firms' lifetime. The further away from the starting value

V_0^E, the lower is the density of V^E because longer and more heterogeneous life paths underlie such values.

Subplot (c) shows the distribution of repayments. It follows directly from the distribution $\Psi(V^E)$ (because V^E is the underlying state variable). Depending heavily on the current period productivity realization, the levels of repayments are separated into two groups. This means, the repayments exhibit two distinct sub-distributions because the difference in repayments of high and low state are relatively large (compare $m_h(V^E)$ and $m_l(V^E)$ in Figure 1.4).

Subplot (d) shows the distribution of bank loans. It also follows directly from the distribution $\Psi(V^E)$ (because V^E is the underlying state variable). It captures the firm size distribution measured by the levels of bank loans. From Figure 1.4 follow that for many V^E the optimal level of banks loans lies around the value $b(V^E) \approx 1.12$ (see relatively flat part in Figure 1.4). This means, many firms get such levels of banks loans so that the mode of the distribution of bank loans lies around this value. Thus, the negative skewness in the distribution of b is the result of the less strongly increasing part of $b(V^E)$ seen in Figure 1.4.

1.4.4 Firm dynamics

By considering now firm distribution of different cohorts separately (i.e., all entrepreneurs of the same age τ), the model allows us to get firm dynamics: Average firms' size, growth and variance of growth at different ages.

First, using the simulation of life path of entrepreneurs in Section 1.4.2 we generate the distribution of promised values $\Psi_\tau(V^E)$ of entrepreneurs at different ages.[22] The development of $\Psi_\tau(V^E)$ for selected cohorts with age $\tau = \{1, 2, 4, 8, 16, 32, 64, 209\}$ is shown in Figure 1.7.

The newborns $\tau = 1$ are all identical with the same starting promised value $V_0^E = -8.36$. Surviving firms then experience either high or low productivity realizations and are updated with higher or lower future promised value levels, respectively. Over time as τ gets larger, histories of productivity realizations get more heterogeneous due to the i.i.d. shocks. The distribution of promised values, $\Psi_\tau(V^E)$, gets more dispersed. In addition, as age advances cohort size becomes smaller because firms have been exiting with the exogenous death rate $1 - \Delta$. Eventually, (almost) all firms of a given cohort exit the market so that the distribution $\Psi_\tau(V^E)$ of old cohorts consist of very few individual observations.

[22]This maps directly into the distributions of bank loans and repayments. The corresponding distributions of b and m are shown in Figure A.7 and A.8 in Appendix A.6, respectively.

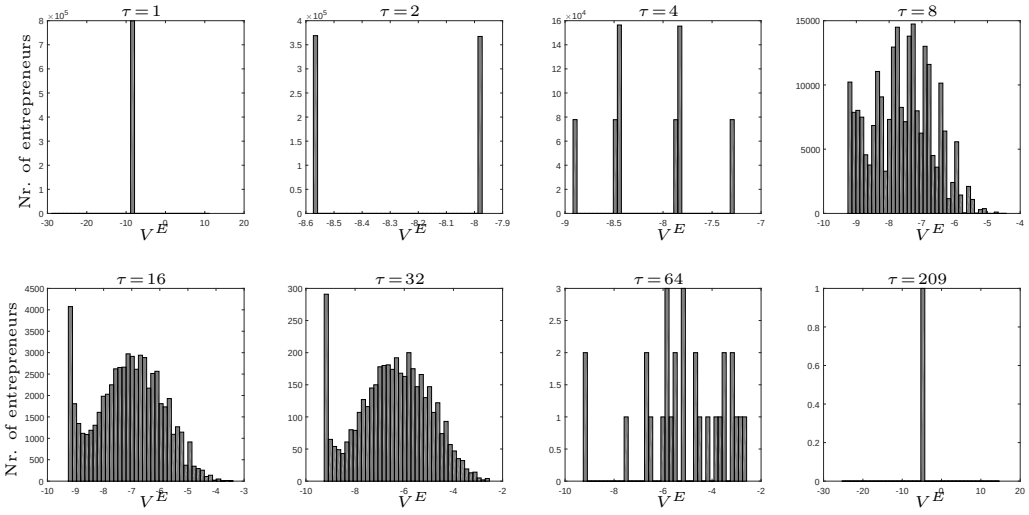

Figure 1.7: Development of entrepreneurs' promised value distributions

Following the cohort distribution, $\left\{\Psi_\tau(V^E)\right\}_{\tau=0}^{\infty}$, we can get firm dynamics such as average size, growth and variance of growth at different ages τ of entrepreneurs. Such firm dynamics are shown in Figure 1.8.[23]

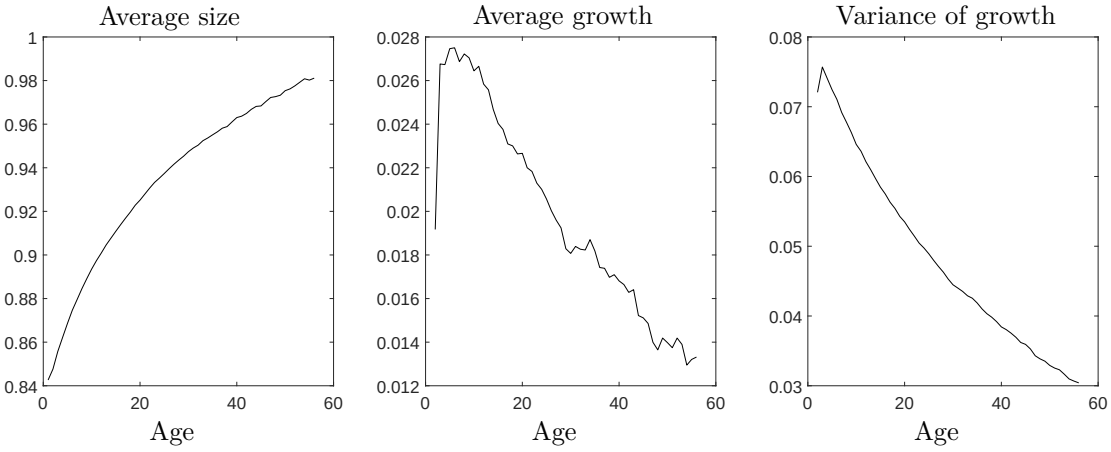

Figure 1.8: Firm dynamics

Figure 1.8 shows in Subplot 1 an increasing average size of firms at different ages.[24]

[23]To see the trend of the firm dynamics more clearly, we plot the 5-year moving average (e.g., the value of the average size at age 10 is the weighted average size of firms with age 10-14.)

[24]There is a decrease in firm size between the one-year-olds and the two-year-olds. To see why, first, notice that the starting promised value, $V_0^E = -8.36$, is at the right end of the steep part of the $b(V^E)$-function; a low productivity shock lowers b more than a high productivity shock increases b. In addition, since the history of productivity shocks is not very heterogeneous after one period (i.e., 50% are high and 50% are low), the decrease from the low productivity shock is directly reflected in the average size. For more periods the history of productivity shocks of entrepreneurs becomes more heterogeneous and the

Firm size is measured in terms of the level of banks loans.[25] Hence, our model predicts a positive relation between firm size and their age, which is in line with empirical observations.

In Subplot 2 we plot firms' average growth rates at different ages. We define the growth rate of a firm at age τ by the percentage change in bank loans relative to last period's loan, $g_\tau \equiv \frac{b_\tau - b_{\tau-1}}{b_{\tau-1}}$, where b_τ and $b_{\tau-1}$ are bank loans of today and of yesterday, respectively. The average growth rate of all firms at age τ is measured by the mean of g_τ among all entrepreneurs in this cohort. The graph shows that firms' average growth is positive, but the rate decreases with firm age.[26] The same holds for the variance of the growth rate (i.e., the variance of g_τ) which is shown in Subplot 3. This means that on average older firms grow less, but in a more stable way.

These patterns are also found by Clementi and Hopenhayn (2006), Gross and Verani (2013) and Verani (2015), and are observed in industry data (e.g., Evans (1987)). This suggests that empirical firm dynamics can be explained by the design of the optimal financial contracts with endogenous borrowing constraints.

1.5 Model applications

In this section, we propose two applications of the benchmark model. First, we study the impact of an increasing production volatility on the equilibrium variables, the aggregate variables and firm dynamics.[27] More specifically, we analyze how a mean-preserving-spread of the firms' productivity influences the equilibrium outcome. Second, we extend the model in a parsimonious form to analyze the macroeconomic consequences of bank regulation. In particular, we show the impact of imposing higher reserve ratios on the equilibrium variables, credit availability at firm level, and the resulting firm dynamics.

---

average is thus less dependent on the level of bank loans corresponding to a specific history of productivity realizations.

[25] Firm size can be equivalently measured by the level of capital employment or labor employment. This can be seen from the linear relation between b and k^*, b and l^*, defined in (1.33).

[26] The observation discussed in footnote 24 is the reason for the outlier of the average growth (and also of the variance) in the first year. Note that the less smooth pattern for young firms comes from the fact that at the beginning firms have less different productivity paths, so that we have in this sense not enough cases of observations. The less smooth pattern for older firm arises since firms are dying and not many observations are left.

[27] The equilibrium variables of concern are always factor prices and the equilibrium share of entrepreneurs.

1.5.1 Production volatility

We set the productivity in high, h, and in low, l, state, respectively, as follows:

$$\theta_h = \mathbb{E}(\theta) + \sigma \tag{1.37}$$

$$\theta_l = \mathbb{E}(\theta) - \sigma, \tag{1.38}$$

where $\mathbb{E}(\theta)$ is the expectation of productivity and is normalized to 1. Apparently, a mean-preserving-spread of the firms' productivity implies an increase in σ.

Keeping all other parameters as given in Table 1.1, we calculate the model equilibrium numerically for different values of σ. The resulting equilibrium parameters are summarized in Table 1.4.[28]

Table 1.4: Comparative statics of σ on equilibrium variables

	$\sigma = 0.15$	$\sigma = 0.25$	$\sigma = 0.3375$	$\sigma = 0.5$	Sign
Interest rate, r	0.04187	0.04170	0.04159	0.04156	-
Wage rate, w	0.161	0.160	0.159	0.157	-
Share of entrepreneurs, λ	0.072	0.076	0.077	0.073	+/-

In an economy with higher production volatility, interest rate and wage rate are monotonically lower. However, the share of entrepreneurs displays an inverse U-shape relationship with production volatility: For volatility below a threshold (approximately $\bar{\sigma} = 0.3375$), the share of entrepreneurs increases as the volatility increases. Whereas for volatility above the threshold, the share of entrepreneurs decreases.

The change in interest rate and the change in the share of entrepreneurs for production volatility above the threshold, $\bar{\sigma}$, are in contrast to Smith and Wang (2006), where firms' capital and labor demand are exogenously given.

1.5.1.1 Impact on firm dynamics

Given the equilibrium factor prices, we can analyze the impact of a rising productivity volatility on firm dynamics as discussed in Section 1.4.4 (i.e., size, average growth, and variance of growth). The underlying mechanism that determines the credit availability at firm level can be attributed to two aspects: Changes in the optimal contract and changes in the distribution of the population in the economy (in terms of promised values).

Volatility effect and equilibrium price effect

[28] $+/-$ means first increase and then decrease, and $-/+$ indicates the opposite.

An increase in the production volatility has two counteracting effects on the optimal contract: A direct "volatility effect" (i.e., ceteris paribus, the impact of an increase in volatility on the optimal contract) and an "equilibrium price effect".

For isolating the volatility effect, we first fix the factor prices constant.[29] A mean-preserving spread of production volatility decreases firms' realized output, $\theta_s R(b)$, in low state ($s = l$), and increases it in high state ($s = h$). On the one hand, lower firms' output in low state suppresses the amount of repayment banks could potentially ask for. On the other hand, however, higher output in high state does not necessarily lead to higher repayment. Notice that higher repayment in good state will drive up the gap between the repayments in the two contingencies. To guarantee truth-telling behavior of the firms, the banks need to spread the gap in future promised values, $V_h - V_l$ (see constraint (IC)). Given the concavity of the profit function, $P(V^E)$, this is costly for the banks. Therefore, banks only ask for more repayment in high productivity state if the cost of spreading future promised values can be compensated. Our quantitative results show that a higher production volatility lowers banks' profits. In other words, the net gain from higher repayment in high state (if at all) is not enough to compensate the loss in repayment in low state. In addition, it lowers banks' expected marginal return from granting bank loans, and thus push down the level of bank loans they grant. This is especially the case for firms with low promised values. The residual from production (i.e., consumption) of these firms is low, which indicates a stricter limited liability constraint. A decrease in the level of output may lead to a binding constraint, which drives down the repayment.[30] Overall, the volatility effect lowers credit availability of firms, especially firms with low promised values, and drives down banks' profits. The higher the volatility the stronger the effect.

Now we consider the equilibrium price effect. As listed in Table 1.4, the equilibrium prices decrease as production volatility rises. Lower equilibrium factor prices drive down the cost of running a firm. This increases firms' profits and marginal return of bank loans in the context of a decreasing return to scale technology. Since higher profits imply higher repayments potentially, banks have more incentive to provide bank loans. This effect is stronger for firms at higher level of promised values. To see this, first notice that the optimal bank loans starts from $b = 0$ at $V^E = V_{min}$ regardless of the production volatility, and approaches the efficient size of firms defined in (1.35) as V^E is larger. Since the efficient level is larger in an economy with lower factor prices, given the curvature of the optimal bank loans, $b(V^E)$, the gap between the level of bank loans under different

[29]For example, take the benchmark equilibrium factor prices, $r = 0.0417$ and $w = 0.1599$.

[30]For firms with high promised values, the limited liability constraint is far from binding (e.g., at the promised values when the repayments are negative). Then this volatility effect is very small, even zero.

production volatility, σ, is larger as V^E is away from V_{min}.

In addition, the increase in bank loans boosts firms' production, $R(b)$. As a result, in high state of a more volatile economy, firms' profits, $\theta_h R(b)$, increase significantly due to an increase in both θ_h and $R(b)$. At the same time, the negative impact of a low productivity realization, θ_l, is mitigated by the high production. Therefore, banks can ask for higher repayment in both states, and thus their profits rise. Overall, the equilibrium price effect leads to an expansion of credit for all firms, especially for those with high promised values, and increases banks' profit. Since the factor prices are lower when the production volatility is higher, the equilibrium price effect is stronger in a more volatile economy.

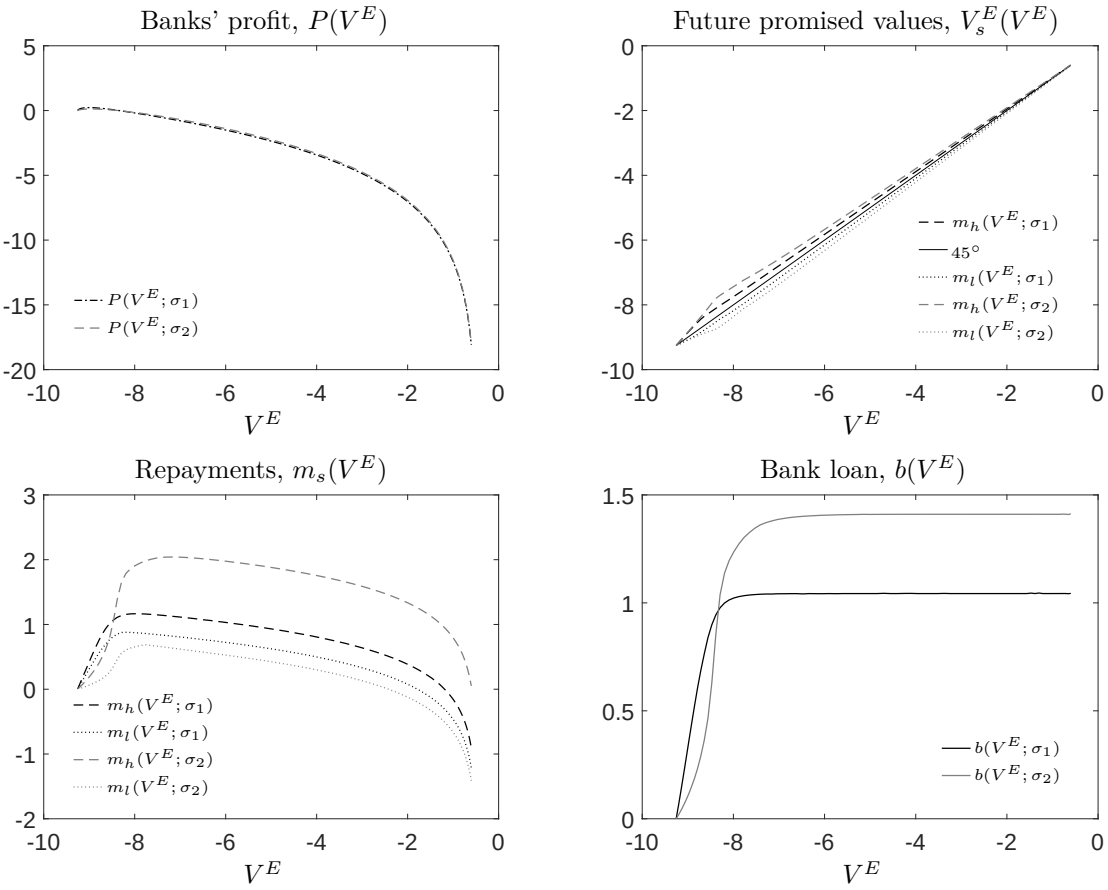

Figure 1.9: Comparative statics of σ on the optimal contract in equilibrium

In Figure 1.9 we illustrate the optimal equilibrium contract under $\sigma_1 = 0.15$ and $\sigma_2 = 0.5$.[31] Both the volatility effect and the equilibrium price effect are accounted for in

[31]We illustrate only the two boundary cases of our numerical analysis, so that the small changes can be recognized more easily. The plots of the optimal contract under other intermediate values of σ lie in between the two cases we showed.

the figure. For firms with low level of promised values (V^E close to V_{min}), the volatility effect dominates. As a result, firms are more financially constrained (i.e., b is lower), banks ask for lower repayments, $m_s, s \in \{h, l\}$, and the banks' profit, $P(V^E)$, is thus lower. In contrast, for firms with higher promised values, the equilibrium price effect dominates. Consequently, firms get more bank loans (i.e., b is larger), firms' profits are higher, which increases repayments in high state, m_h, and mitigate the negative impact of productivity shock in low state on repayments (the mitigation can be seen from the fact that $m_l(\sigma_2)$ is not much lower than $m_l(\sigma_1)$, where $\sigma_2 > \sigma_1$). In the end, to induce truthful behaviors of firms, the gap between future promised values is larger in a more volatile economy (i.e., $|V_h(\sigma_2) - V_l(\sigma_2)| > |V_h(\sigma_1) - V_l(\sigma_1)|$).

Distribution effect

So far we have discussed the impact of an increasing production volatility on the optimal contract. To understand the underlying mechanism that determines credit availability at firm level, we still need to see the changes in the distribution of promised values for entrepreneurs at different ages, τ, in the economy, $\Psi_\tau(V^E; \sigma)$ defined in 1.3.1.2.

Entrepreneurs' promised values start from $V^W(0; \sigma)$ under indifferent occupational choice (1.26), and grow on average as they get older.[32] In an economy with higher production volatility, equilibrium factor prices are lower. Thus, the lifetime expected utility of new-born workers, $V^W(0; \sigma)$, is lower, and so is that of new entrepreneurs. Given the monotonically increasing relation between promised values and the level of bank loans, new firms are more financially constrained in a volatile economy. Moreover, the lower initial promised value is propagated over firms' lifetime (see section 1.3.1.2), and tends to lower the level of bank loans to all firms. Furthermore, the spread in future promised values in the two contingencies is larger as the production volatility increases, the distribution of promised values is more dispersed. This increases the growth rate of firms' bank loans, and the variance of growth.

We summarize the impact of an increasing production volatility on credit availability at firm level in Table 1.5. Specifically, we calculate at four σ values, the average firm size (in terms of level of bank loans granted to firms), average growth rate and variance of growth. Furthermore, to see more clearly how production volatility influences firms of different ages differently, we decompose the firms into three age groups, 1-20, 21-40, and 41-60.[33]

[32]Notice that the lifetime expected utility of a worker is actually $V^W(0; r^*(\sigma), w^*(\sigma))$, where $\{r^*(\sigma), w^*(\sigma)\}$ are the equilibrium factor prices under production volatility, σ. We write $V^W(0; \sigma)$ for simplicity.

[33]We calculated the values for smaller age groups (e.g., groups of every 10 years, 1-10,11-21, etc.) and for larger age range (e.g., age groups until the cohort of 100 years old, which accounts for 99.98% of the population), the results are robust.

Table 1.5: Comparative statics of σ on firm dynamics

	$\sigma = 0.15$	$\sigma = 0.25$	$\sigma = 0.3375$	$\sigma = 0.5$	Sign
Average firm size, b	0.958	0.884	0.863	0.901	-/+
Age group 1-20	0.953	0.869	0.843	0.882	-/+
Age group 21-40	0.977	0.938	0.939	0.973	-/+
Age group 41-60	0.993	0.970	0.979	1.020	-/+
Average firm growth	0.004	0.020	0.040	0.086	+
Variance of growth	0.01	0.06	0.12	0.31	+

The average firm size first decreases and then increases as production volatility increases. This indicates a dominating volatility effect and distribution effect at low production volatility, both of which tend to reduce the level of bank loans. And then a strong equilibrium price effect at high production volatility which drives up the average bank loans. Moreover, both the average firm growth and variance of growth increase monotonically with production volatility, as a result of a strong equilibrium price effect and a distribution effect that disperse the promised values.

Furthermore, higher volatility leads to a severe credit rationing among young firms (age 1-20). This comes from a strong volatility effect. Even though the equilibrium price effect compensates the volatility effect at $\sigma = 0.5$, the level of bank loans is still 7.5% lower than firms of the same group under $\sigma = 0.15$. And at $\sigma = 0.3375$ where the price effect is less strong, the decrease in bank loans is 11.5%. As firms grow older the equilibrium price effect becomes stronger, and thus firms are less financially constraint under high volatility. This is reflected from a flatter change of firm size as volatility increases, indicating the volatility effect is offset or even dominated (firms of age 41-60 get 2.7% more loans under $\sigma = 0.5$ than under $\sigma = 0.15$, and the level of bank loans under $\sigma = 0.3375$ is only 1.4% lower than under $\sigma = 0.15$).

In sum, for a relatively large range of production volatility ($\sigma < 0.3375$), higher volatility lowers the average firm size in the economy. Young firms are the most vulnerable ones financially to productivity shocks. When production volatility is very strong, the young firms are still the most financially constrained, but the average firm size is larger. Yet, one should notice that the share of entrepreneurs decreases with higher volatility. Furthermore, there is a tradeoff between average growth and variance of growth: Firms grow faster in an economy with higher productivity volatility. However, the higher growth comes at the expense of a higher variance of growth.

In Figure A.9 in appendix A.6, we illustrate the development of average firm size, average growth rate and variance of growth at all ages (1-60). The quantitative results confirm our conclusions above.

1.5.1.2 Impact on aggregate variables

An increase in production volatility induces the following changes on aggregate variables. Both total labor supply and capital supply decrease as a result of lower equilibrium factor prices. Aggregate bank loans, B, first decrease and then increase as the volatility increases. Notice that the value of aggregate variables are equal to the corresponding average value, because the size of the agents is normalized to 1 (see (1.19)-(1.22)). Thus, the underlying mechanisms that drive the change in aggregate bank loans are the same as the ones we discussed for average bank loans. Aggregate repayments, M, show the same pattern as the aggregate bank loans, indicating an effective limited liability constraint. Furthermore, according to the linear relationship defined in (1.36), total capital demand, K^D and total labor demand, L^D, both display a similar U-shape pattern.[34] Finally, the ratio of bank equity to bank loans (the bank's equity ratio) increases as production volatility increases, indicating a change in the composition of the banks' balance sheet on the liability side (i.e., a shift towards more equity).[35] The quantitative results are summarized in Table 1.6.

Table 1.6: Comparative statics of σ on aggregate variables

	$\sigma = 0.15$	$\sigma = 0.25$	$\sigma = 0.3375$	$\sigma = 0.5$	Sign
Total labor supply, L^S	0.289	0.288	0.287	0.284	-
Total capital supply, K^S	0.170	0.162	0.156	0.155	-
Total bank loans, B	0.96	0.88	0.86	0.90	-/+
Total labor demand, L^D	3.76	3.49	3.43	3.62	-/+
Total capital demand, K^D	2.49	2.30	2.25	2.35	-/+
Total repayments, M	0.95	0.87	0.85	0.89	-/+
Equity ratio, E/K^D	0.108	0.147	0.164	0.165	+

1.5.2 Bank regulation: Reserve ratio

Suppose that banks must hold a share μ of the deposits as reserves. This implies that apart from banks' equity, E, only a share $1-\mu$ of the deposits, $(1-\lambda)D$, can be supplied

[34] According to (1.36), both level of bank loans, B, and the equilibrium factor prices, $\{r, w\}$, have an impact on the total capital and labor demand, K^D and L^D. The intuition is straightforward: As more credit becomes available, firms employ more capital and labor to increase profits. At the same time, if factor prices decrease, firms increase their demand for the corresponding input factor. However, from the values of K^D and L^D at different σ-levels listed in Table 1.6, the former channel dominates.

[35] Intuitively, this is mainly due to the change in the distribution of firms' promised values. As is discussed in Appendix A.5, banks accumulate equity from firms with low promised values (i.e., expected repayment is higher than bank loan) and the opposite occurs for firms with high promised values. Since the initial promised value, $V^W(0;\sigma)$ decreases as productivity volatility increases, the distribution of firms' promised values tends to shift towards lower promised values, which leads to an increase in banks equity.

as capital to finance firm production. Therefore, the asset market clearing condition is now given by

$$\lambda K \leq (1-\lambda)(1-\mu)D + \lambda E. \tag{1.39}$$

Using the parameter values given in Table 1.1 we calculate the equilibrium numerically for $\mu = \{0, 0.2, 0.4\}$. The outcomes for $\mu = 0$ coincide with the benchmark case in Section 1.4. The quantitative results are summarized in Table 1.7. In an economy with higher reserve ratio, the equilibrium interest rate is higher, wage rate and the share of entrepreneurs are lower.

Table 1.7: Comparative statics of μ on equilibrium variables

	$\mu = 0$	$\mu = 0.2$	$\mu = 0.4$	Sign
Interest rate, r	0.0417	0.0424	0.0436	+
Wage rate, w	0.1599	0.1594	0.1587	-
Share of entrepreneurs, λ	0.0762	0.0757	0.0748	-

1.5.2.1 Impact on firm dynamics

What is the impact of a rising reserve ratio on firm dynamics and the credit availability at firm level? Notice that a change in reserve ratio, μ, has no direct impact on the optimal decisions at individual level (saving and labor supply decision by workers, labor and capital employment by entrepreneurs and the optimal financial contract). Only through an equilibrium price effect on the optimal contract and the resulting distribution changes.

Equilibrium price effect

The changes in interest rate and wage rate have counteracting effects on the optimal contract. Our quantitative results show that as the reserve ratio rises, the efficient size of firms (measured by the level of bank loans), defined in (1.35), decreases. The tighter reserve ratio raises the interest rate, and thus depresses availability of credit in the market.

Distribution effect

In addition, the initial promised values of firms decreases as the reserve ratio increases.[36] As a result, in particular new firms are more financially constrained. Furthermore, firms at older ages have higher average promised values in an economy with high reserve ratios. This tends to increase the credit availability of the old firms.

In Table 1.8 we summarize the impact of an increasing reserve ratio on average firm size, average growth and variance of growth. Again, to see more clearly how firms of

[36]The values of V_0^E at $\mu = \{0, 0.2, 0.4\}$ are -8.3549, -8.3553 and -8.3554, respectively.

Table 1.8: Comparative statics of μ on firm dynamics

	$\mu = 0$	$\mu = 0.2$	$\mu = 0.4$	Sign
Average firm size, b	0.8836	0.8830	0.8832	-/+
Age group 1-20	0.8693	0.8681	0.8674	-
Age group 21-40	0.9381	0.9400	0.9439	+
Age group 41-60	0.9698	0.9727	0.9780	+
Average firm growth	0.01992	0.01988	0.01983	-
Variance of growth	0.058	0.057	0.055	-

different ages are influenced differently, we decompose the population of firms into three age groups, 1-20, 21-40, 41-60.

Average firm size first decreases and then increases. The increase comes from a strong positive impact of an increasing average promised values on credit availability as the reserve ratio becomes very large ($\mu = 0.4$). This effect is strong enough to overcome not only the propagation effect of a lower initial promised value, but also the negative equilibrium price effect on the optimal bank loan, $b(V^E)$.[37] As we said, this positive effect mainly influences the credit availability of older firms. This is confirmed by looking at the firm size of different age groups: Only for firms older than 20 years, banks' credit expands, whereas for young firms credit shrinks as a result of the negative equilibrium price effect and the lower initial promised value. Furthermore, as the reserve ratio increases, firm growth decreases, and its variance decreases as well. This is because of a dominating equilibrium price effect: Since the efficient firm size is smaller as the reserve ratio rises, the range of bank loans decreases, and thus also the average growth and variance of growth.

In sum, except for unrealistic high levels of reserve ratios, firms operate in smaller size in an economy with higher a reserve ratio. In addition, there is a redistribution of credit from young firms towards older firms, implying a worse-off financial situation of the young firms. In the end, firms grow more slowly but more steadily as the reserve ratios increase. The lower growth rate of firms is accompanied by less volatile growth.

In Figure A.10 in appendix A.6, we illustrate the development of average firm size, average growth rate and variance of growth at all ages (1-60). The quantitative results confirm our conclusions above.

[37] We verify numerically that it is indeed due to an increasing average promised values that drives up the average firm size. Specifically, we want to isolate the effect on average firm size due to the transition in promised values with the other two negative effects. To do this, we use the distribution of promised values calculated with transition function $V^E(\theta_s, V^E)$ under $\mu = 0$, in combination with the optimal bank loan $b(V^E)$ and the initial promised value $V^W(0)$ under $\mu = 0.4$, to get average bank loans. The value is smaller than the average bank loans under $\mu = 0$. This implies that the gap between this value and the actual average bank loan under $\mu = 0.4$ is due to a strong impact of increasing average promised values on the level of bank loans.

1.5.2.2 Impact on aggregate variables

In Table 1.9, we summarize the impact of an increase in the reserve ratio on aggregate variables. Again, because the size of the agents is normalized to 1, the aggregate value is the corresponding average value (see (1.19)-(1.22)).

Table 1.9: Comparative statics of μ on equilibrium variables

Total labor supply, L^S	0.288	0.287	0.284	-
Total capital supply, K^S	0.162	0.197	0.253	+
Total bank loans, B	0.8836	0.8830	0.8832	-/+
Total labor demand, L^D	3.491	3.497	3.514	+
Total capital demand, K^D	2.297	2.284	2.265	-
Total repayments, M	0.870	0.868	0.867	-
Equity ratio, E/K^D	0.147	0.153	0.164	+

Total labor supply decreases and total capital supply increases, due to the decline of the wage rate and the rise in the interest rate. Total bank loans first decrease and then increase for the same reason as the average bank loans. However, despite the linear relationship defined in (1.36), total labor demand increases, whereas total capital demand decreases. This results from substitution effects between input factors at firm level. Total repayments decrease monotonically, because the increase in total bank loans mainly comes from an increase in the level of bank loans to older firms, for which the limited liability constraint is no longer tight. In the end, the banks' equity ratio increases in an economy with higher reserve ratio. To see this, notice that as the reserve ratio increases, for each unit of deposit a bank acquires, the share that can be given as loan decreases. Therefore, profit maximizing banks shift towards other sources to finance bank loans. Since banks' equity is the only other source on the liability side, banks shift to more equity.[38]

1.6 Discussion of dynamic programming

In this section, we discuss problems one encounters in solving the dynamic contract. They are, among others, starting value problems, extrapolation issues, sensitivities to functional forms and to parameter values, and issues related to the simulation.

[38]Since the initial promised value $V^W(0;\mu)$ decreases as reserve ratio μ increases, the building-up of bank equity follows the same intuition as discussed in footnote 35.

1.6.1 Starting value problems

Dynamic programming problems are sensitive to initial guesses of value function and policy functions. In general, the convergence of dynamic programming algorithms is limited to a region close enough to the solution. This issue is especially obvious in the dynamic contract problem with relatively unconventional constraints. Therefore, our problem requires proper guesses of the starting values. We tried two ways: First, educated guess derived from the functional form of the utility function and, second, a "ground search". For the educated guess, we consider a contract under perfect information with constant consumption for all periods and states. Then the value function, $P(V^E; r, w)$, which can serve as the starting value, is an affine linear transformation of the inverse of the utility function. In the ground search, we programmed a loop over a broad grid set of $\{b(V^E),\, m_h(V^E), m_l(V^E),\, V_h^E(V^E),\, V_l^E(V^E)\}$. We calculated for all combinations of the grid points the corresponding bank's profits and then checked which of the combinations of the grid points maximize banks' profits given that it fulfills all the constraints. These grid points are supposed to be somewhere in the region close to the solution of the optimal contract and can thus serve as the starting values. Overall, the initial guesses from the two ways are both good enough for solving the dynamic contract in our model. In the end, we used the first way to get the initial guesses as the second way requires relatively long computation time and a large amount of storing memory.[39]

1.6.2 Extrapolation errors

In the numerical algorithm we generate a finite number of Chebychev grid points on the interval of the state variable.[40] Chebychev grid points have superior performances in function iterations in dynamic programming, yet, an extrapolation problem arises: The interval on which value function and policy functions are defined is larger than the range of the grid points, and thus extrapolation may be needed. With cubic splines interpolation, extrapolation close to the surroundings of the two grid boundaries is embedded in the code and thus performed automatically. However, the default extrapolation cannot guarantee that the image of the policy function remains in the domain of the state variable. To prevent this, we manually replaced the lowest Chebyshev point with the lower bound of the interval of the state variable before using it.

---

[39] Note that even with only four grid points for each of the five choice variables there are already $4^5 = 1024$ combinations to be calculated and checked.

[40] For Chebychev grid points we follow Judd (1998): The m grid points $\{x_k\}_{k=\{1,\ldots,m\}}$ are set according to the coefficients of the Chebychev polynomial. We compute m Chebychev interpolation points $z_k = -\cos\left(\frac{2k-1}{2m}\pi\right)$ on $[-1, 1]$. Then we adjust it to our interval $[a, b]$, such that $x_k = (z_k + 1)(\frac{b-a}{2}) + a$.

1.6.3 Sensitivity to parameter values and functional forms

Given the complexity of the dynamic contract problem, the numerical outcomes and the convergence of the iterations are sensitive to functional forms and parameter values. In our case, for example, with log-utility we would see a U-shaped $b(V^E)$, which may not necessarily lead to the same firm dynamics as we observe in Figure 1.8. Further, convergence of the problem is sensitive to the combinations of parameters. For example, not all combinations of (r, w) may guarantee convergence of the value function iteration when calculating the optimal dynamic contract. However, for combinations of (r, w) close to the equilibrium solution, the dynamic contract problem is in general stable; it always results in proper optimal contracts.

1.6.4 Simulation issues

To get the equilibrium, we grid-search (r, w)-combinations manually and check for the equilibrium conditions to hold. To have monotonicity in the aggregation variables of the simulation and comparability of different outcomes from the grid search, it is important that the simulation reflects a stationary distribution of life paths for all compared (r, w)-combinations. Otherwise, one cannot identify whether changes in the zero-profit condition and the capital market clearing condition come from the effect of updated (r', w') or from a changed combination of life paths. Furthermore, as is described in more detail in Appendix A.3.4 and A.4, we use the observable fact that banks' profit and the excess demand in the capital market are both decreasing in r and w. In addition, the gap between the banks' profit and the excess demand in the capital market is decreasing in r and increasing in w. We use these signs and the (at least locally) observable monotonicity of the two conditions to restrict the region where the optimal equilibrium lies and get the direction for further searching.

1.7 Conclusion

This paper adds to the literature by modeling a dynamic credit relationship between banks and entrepreneurs in a general equilibrium model – which determines simultaneously the wage and interest rate and the share of entrepreneurs, and delivers firm dynamics. We have households who decide, at the beginning of their life, to become either a worker or an entrepreneur. Workers supply labor and save in the form of annuity deposits. Entrepreneurs run firms and employ labor and capital for production. Productivity is stochastic and private knowledge to the entrepreneur. The production costs are financed with bank loans. To overcome the information asymmetry, loans and repayments are

determined in long-term financial contracts with banks. More specifically, the financial contract between banks and entrepreneurs derived from a recursive formulation determines the optimal level of bank loan, state-contingent repayments and future promised values given today's promised values. The contracts are promise keeping and incentive compatible and fulfill limited liability and credibility constraints. In equilibrium banks make zero profit from the contract and the labor, the capital and the goods markets are cleared. The general equilibrium structure allows determining the wage, the interest rate and the share of entrepreneurs in equilibrium, as well as the size distribution of firms. Further, we get firm dynamics arising through the optimal financial contracts: The size of firms, measured by their level of bank loans, increases with the age of firms while their average growth and the variance of growth decreases with age.

2 Explaining structural change towards and within the financial sector[1]

Joint with Josef Falkinger and Sabrina Studer

2.1 Introduction

Financialization and inequality are topics that stir up the public debate – among experts as well as outside the scientific community. Discussions about financialization have gained momentum by the financial crisis (Greenwood and Scharfstein, 2013; Philippon and Reshef, 2012, 2013); the inequality debate was brought "in from the cold" (Atkinson, 1997) towards the end of the last century and has reached the center court recently with the Piketty book (Piketty, 2014). This paper argues that the two phenomena are genuinely related to each other. Structural change towards and within the financial sector, as observed over the last three decades, enhances inequality. And rising inequality fosters financialization.

We present our argument in a model that comprises the most basic tools provided by economics for analyzing sectoral structure and distribution. Financialization means two things: The weight of financial business relative to non-financial business increases and the type of financial business changes. From a macroeconomic perspective the first aspect can be summarized as structural change towards the financial sector: The financial sector expands relative to the production sector. We do not approach this question from a monetary or financial aspect like the nominal transaction volume of the financial relative to the real sector. Our perspective is a real economics one: The financial sector employs resources and generates income for the resources employed. The relevant measures are therefore employment and income or output shares; the essential component to be modeled are the production function of the financial sector and the demand function for financial services. For capturing the second aspect of financialization – the shift from conventional banking type activities to sophisticated modern finance – an appropriate model structure requires to have two separate subsectors within the financial sector which differ in their

[1]This chapter is a revised version of the Working Paper No. 206 from the Working paper series / Department of Economics at the University of Zurich.

demand and production characteristics. In sum, we have therefore a three sector model – one production sector and two financial subsectors.

Inequality requires to have heterogeneous agents which differ in their endowments. In our model we have low-skilled and high-skilled workers. They are mobile between sectors and cost-minimal skill-intensities differ across sectors. As a consequence, the interaction between sectoral structure and inequality comes through the skill premium. The focus on inequality between low-skilled and high-skilled workers is on the one side motivated by the empirical fact that the rise in inequality over the last decades has been driven to a large extent by skill premia and skill composition, as the ample evidence from the skill-bias literature shows (for instance, Machin and Van Reenen (1998); Piketty and Saez (2003)). On the other side, we see it as a first important step, which later might be complemented by elements which focus on the functional distribution of income between workers and capitalists or on rents. There is capital in our model; it must be. After all, financial markets have the purpose to transform, under risk, current resources into future production possibilities. This requires, on the one side, saving decisions and, on the other side, capital investment into revenue bearing inputs to future production. In our model, returns on capital are generated by two different types of technologies (robust and risky) which transform savings into future consumption possibilities.

Structural change can be caused by the supply side: Changing endowments or technical change. The huge literature on directed technical change, for instance, has emphasized this channel (Acemoglu, 2002). There is, however, also an important role for the demand side. Although often neglected, income effects are essential for aggregate developments (Boppart, 2014, 2015; Föllmi and Zweimüller, 2008). We account for demand side effects by assuming that agents have quasi-homothetic preferences of the Stone-Geary form. The specific finance aspect enters the demand side of our model through the following channel: Demand for financial services comes from the need to manage portfolios and to finance investments into profitable projects in a way that reflects the preferences of the agents who own the endowments of the economy. Stone-Geary preferences account for the fact that part of the savings is motivated by future subsistence expenditures.

In our model the finance industry correctly assesses risks and productivity of investment projects and earns no rents. This is against popular views; neither does it reflect a common view of the authors of this paper. Actually, there are many sources for imperfections in the financial sector. For instance, prices and payoffs of financial products may be distorted by neglected correlation (Studer, 2015), or insider knowledge and barriers to entry generate rents for financial intermediation. A salient example is the so called finance premium. There is convincing evidence that a finance premium exists (Célérier and Vallée, 2016; Philippon and Reshef, 2007, 2012), that is, the same type of

labor earns more in a finance job than in other occupations. Nonetheless, from a methodological point of view, we consider it as important to start with a benchmark model in which distortions are kept at a minimum. Given the firm basis of such a benchmark, one can then be bold in looking at the role of imperfections which certainly exists in reality in general and in the financial business in particular. Arguably, rents can be more easily extracted when they go along with the tide rather than against it. So it is important to know if outcome changes are supported by changes in economic fundamentals. In a supplementary section we analyze a series of extensions which show how distortions affect the comparative-static results of the benchmark model. Moreover, in the quantitative implementation of our model, we separate the rent component of the expansion of the financial sector, in particular new finance, from the part that is driven by economic fundamentals.

There is a long literature on the impact of financial development on economic growth (Levine, 2005).[2] The causes of financial sector growth and the changing structure of financial activities, which are the topic of this paper, have been less scrutinized. The literature related to our paper in a more narrow sense is rich as far as the empirical side is concerned. In particular, Philippon and his co-authors did pioneering empirical work on financialization. On the theoretical side the situation is quite different. To our knowledge there are only two attempts to explain structural change towards finance in a general equilibrium framework. Philippon (2012) sketches in his notes a 2x2 model with a real and a financial sector both producing with capital and labor. The financial sector produces intermediation services for households and firms. The focus is on the equilibrium effects of changes in intermediation costs. Improvements in financial intermediation tend to raise real wages but have in general an ambiguous effect on the GDP-share of the financial sector. The GDP-share of finance rises if more firms need intermediation services. Structural change between services for safe assets and services for risky investments or wage inequality are not addressed nor do income effects play a role for the relative size of the financial compared to the real sector. There is only one type of labor, one interest bearing asset and preferences are homothetic. Moreover, there are two types of households - infinitely living saver households and households which live two periods and borrow when young. By contrast, in our paper all households live for two periods and save when young; savings can be invested in a portfolio of safe and risky assets. The second theoretical

[2]While the dominant view in this literature was that financial development is positive for growth, a more skeptical view has emerged in the recent past. Gründler and Weitzel (2012) or Law and Singh (2014) provide evidence that more finance is good for growth at low levels of financial development but harmful beyond a certain threshold. Financial sector growth seems to harm in particular skill-intensive (Kneer, 2013) and R&D intensive (Cecchetti and Kharroubi, 2015) industries. Moreover, negative growth effects are robust if different measures of financialization are used, for instance market capitalization rather than credits (Rousseau and Wachtel, 2011) or the employment share of the financial sector (Capelle-Blancard and Labonne, 2011). Beck et al. (2012) find that in particular the shift from enterprise credits to household credits is detrimental for growth and inequality enhancing.

explanation of structural change towards finance is provided by Gennaioli et al. (2014). Like in Philippon (2012) a 2x2 framework is considered and structural change within the financial is not in the focus of the paper. The real sector produces with capital and labor, the financial sector consists of financial intermediation experts in whom investors trust. Therefore they are willing to pay them fees. Like in our set-up households live two periods and save when young. Moreover, they also account for risky assets. Inequality among households, however, plays no role. The saving decision is exogenous - young households save the entire wage - and the portfolio choice is determined by mean-variance preferences. The main driver for structural change towards finance in their model is the idea that financial intermediation services are not only required for the financing of new capital but also for the preservation of the entire stock of capital accumulated over time. Since in a Solow type growth model the capital coefficient increases, the share of financial services in GDP increases, too. In our model, which focuses on comparative-static equilibrium effects of skills and endowments, technologies and preferences, no long-run accumulation effect is considered.

The structure of the paper is as follows. The next section shows the three facts that the paper wants to explain. Section 2.3 outlines the formal structure of our 3x3 model and its building blocks. Section 2.4 analyzes the production equilibrium, Section 2.5 derives the demand for goods and financial services. Section 2.6 summarizes the effects of inequality on the sectoral structure of the economy. In Section 2.7 the general equilibrium is characterized and comparative-static effects are derived analytically. Section 2.8 confronts the theoretical results with empirical evidence from the U.S.. Moreover, a numerical exercise is provided. In the appendix we provide supplementary material on extensions and model variants to address the effects of distortions and to examine the robustness of our results. Main conclusions are summarized in the last section.

2.2 Facts to be explained

Three facts motivate our analysis: The rising weight of the financial sector in total economic activity, the rising weight of new finance activities within the financial sector and the rise of inequality measured by the wage premium of skilled labor. The following figures show the development for the U.S. economy over the period from 1980 to 2013 (based on the Current Population Survey data).[3]

In Figure 2.1 the weight of the financial sector is measured by the wage ratio Ψ, that is the sum of wages earned in the financial sector divided by the wage sum earned in

[3]Data from IPUMS-CPS by King et al. (2010). Survey years 1980-2013 represent years 1979-2012 because households are surveyed about last year's job. This means whenever we talk about a year the data considered represent the situation a year before.

the other sectors of the economy.[4] We show two measures: The actual finance ratio, Ψ_{actual}, and the normalized one, $\Psi_{normalized}$. As emphasized in the introduction, actual wage earnings in the financial sector comprise a substantial finance premium. Moreover, working hours in the finance industry are higher than in the other sectors of the economy. In our normalized measure, we adjust for these factors and calculate the finance ratio by assuming that employees in the financial sector work equal hours and earn the same wage as the workers (with comparable skills) employed outside the financial sector.[5]

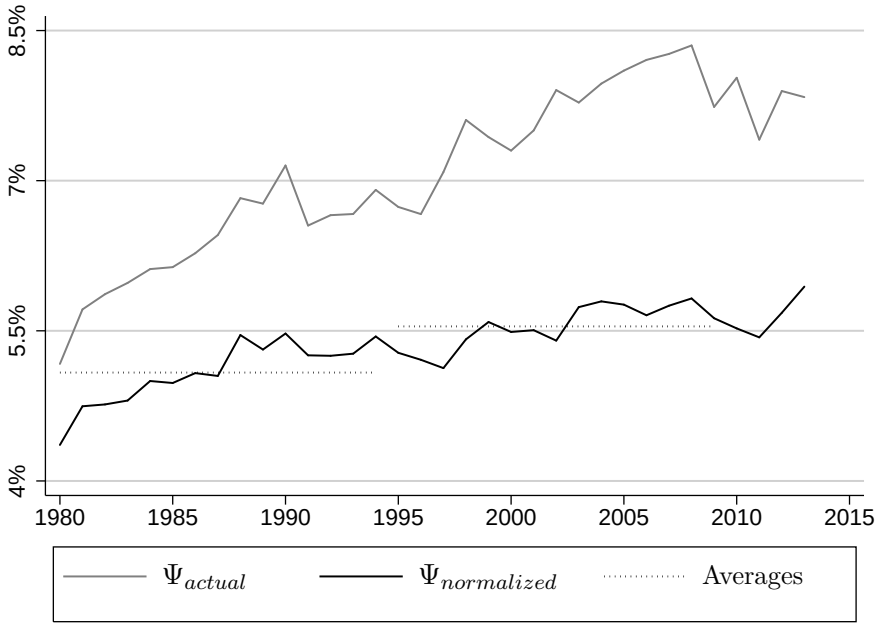

Figure 2.1: Wage sum ratios of the financial sector

Notes: Ψ measures the ratio of the total wage sum in finance vs. the rest of the U.S. economy. "Actual" uses the observed sector-specific hours worked and hourly wages (for low-and high-skilled), whereas "normalized" uses the X-sector hours worked and hourly wages (for low-and high-skilled). Survey years from 1980-2013. Averages of normalized ratio for periods 1980-1994 and 1995-2009, respectively. Source: Own calculations based on CPS.

Figure 2.2 shows the wage ratio of new finance, Φ, measured by the sum of wages earned by workers employed in the new finance activities divided by the wages sum earned in the traditional financial sector. The traditional financial sector comprises banking, credit agencies and insurance; new finance activities consist of security and commodity brokerage and investment companies. Again, a normalized new finance ratio $\Phi_{normalized}$ is shown beside the actual ratio Φ_{actual}, where working hours and wage rates from outside the financial sector are used to calculate the normalized measure.

[4]Employment or value added ratios too would show the increased weight of the financial sector and in particular of new finance activities.

[5]Section 2.8 describes data and measures in more detail.

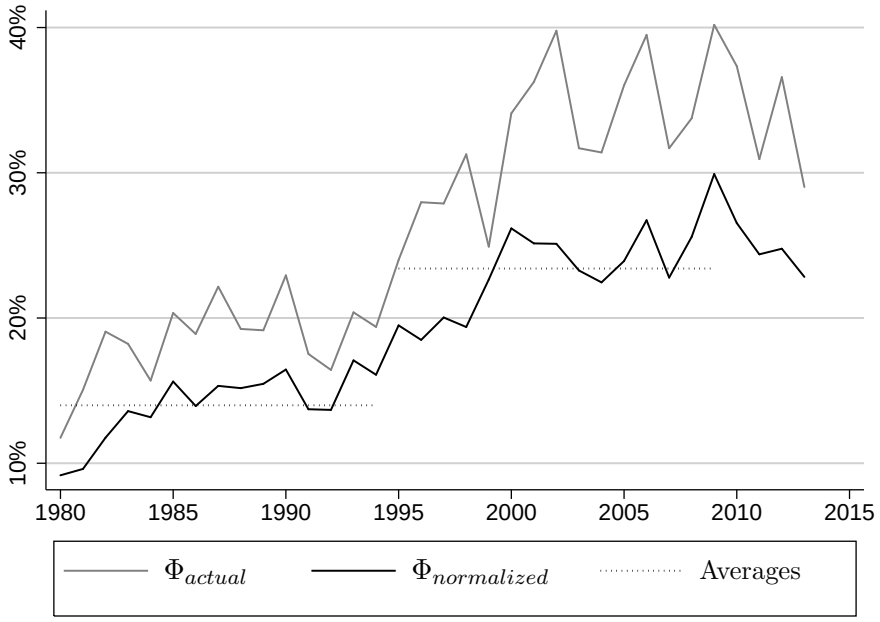

Figure 2.2: Wage sum ratios within the financial sector

Notes: Φ measures the ratio of the total wage sum in "new finance" vs. "traditional finance". "Actual" uses the sector-specific hours worked and hourly wages (for low-and high-skilled), whereas "normalized" uses the X-sector hours worked and hourly wages (for low-and high-skilled). Survey years from 1980-2013. Averages of normalized ratio for periods 1980-1994 and 1995-2009, respectively. Source: Own calculations based on CPS.

Since our baseline analysis focuses on the economic mechanisms in a perfect equilibrium framework, it is only appropriate for explaining the development revealed by the normalized measures of financialization. Possible additional drivers of financialization which could explain the gap between normalized and actual weights of finance and new finance are addressed in the extensions analyzed in Appendix B.2. Yet, as Figure 2.1 and Figure 2.2 show, even the normalized measures clearly reveal a twofold structural change towards and within the financial sector. The normalized wage ratio of the financial sector, $\Psi_{normalized}$, increased from 4.36% in 1980 to 5.94% in 2013 and the normalized wage ratio of new finance, $\Phi_{normalized}$, rose from 9.17% in 1980 to 22.83% in 2013. For the quantitative implementation of our model and the comparative-static equilibrium results we exclude the post-crisis years and compare average values for the period 1980-1994 with the respective average values for 1995-2009. The average normalized wage ratio of finance in the total economy was 5.08% in the period 1980-1994 and increased by 9% to 5.54% in the period 1995-2009. The average normalized wage ratio of new finance increased by 67% from 13.99% in the period 1980-1994 to 23.41% in the period 1995-2009.

The twofold structural change shown in Figure 2.1 and Figure 2.2 has been accompanied by a strong rise in inequality, including wage inequality in particular. Figure 2.3 shows the third fact that we want to explain by our analysis – the rise of the wage pre-

mium ω earned by skilled workers compared to the wage of the unskilled. As shown in Figure 2.3 this skill premium (measured by the normalized ratio of the hourly wage of a skilled worker outside the financial sector divided by the hourly wage earned by unskilled labor) increased from 1.55 in 1980 to 1.91 in 2013. Comparing averages for the two periods 1980-1994 and 1995-2009 the increase of the skill premium we have to explain by our analysis is 14% from 1.62 to 1.85.

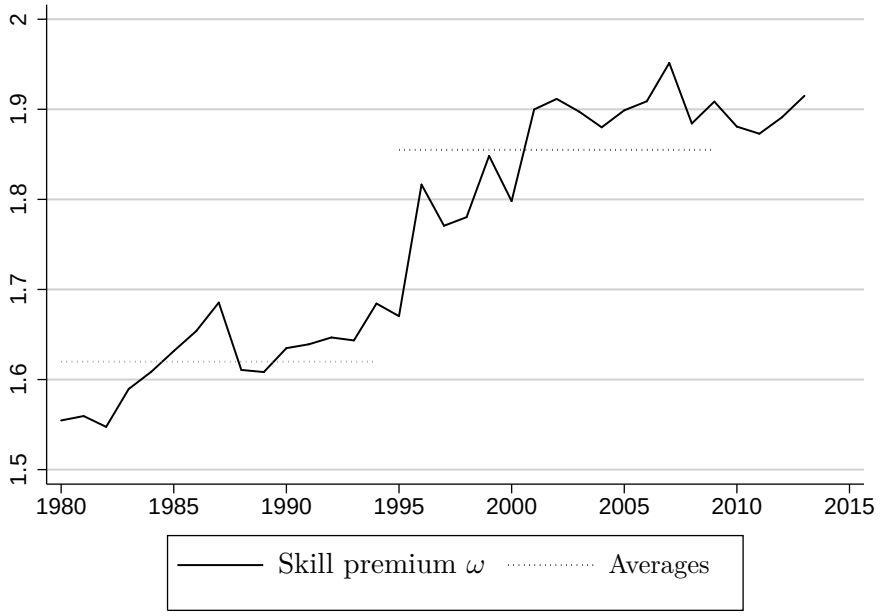

Figure 2.3: Skill premium

Notes: ω measures the "normalized" skill premium (i.e., hourly wage of high-skilled labor in X-sector divided by hourly-wage of low-skilled labor in X-sector). Survey years from 1980-2013. Source: Own calculations based on CPS.

2.3 Model

2.3.1 Model set-up

We model a 3 sector, 3 factor economy. There is a production sector X and a finance sector Z with two subsectors Z_1 and Z_2. All sectors employ low-skilled and high-skilled workers. Produced goods are used for consumption and investment. For transforming savings into future consumption possibilities, more or less risky technologies are available which use capital as input and deliver consumption goods as output in the next period. (As an extension we present a variant of the model, in which capital is used in the X sector to set up firms.) Financial services have the function to support the transformation of savings into future consumption possibilities. Services Z_1 are used for safe savings.

Services Z_2 provide state-dependent instruments and are used for savings in securities with risky returns.

We consider a (static) two-period OLG economy. The future $t = 1$ is uncertain. It consists of a set Θ of distinguishable events and a set $\bar{\Theta}$ of events which are indistinguishable in $t = 0$. The future state space is $\left\{\{\theta|\theta \in \Theta\}, \bar{\Theta}\right\}$. We have prob($\Theta$)=$\mu$ and prob($\theta|\Theta$)=π_θ with $\sum_{\theta\in\Theta} \pi_\theta = 1$.[6] For $\theta \in \Theta$, state-contingent investment possibilities are available which pay off if and only if state θ is realized. No state-contingent investment possibilities exist for $\bar{\Theta}$ which reflects "true uncertainty".

2.3.2 Saving decision and portfolio choice

There are N agents who live for two periods. They are endowed with a skill level and work as either high-skilled or low-skilled worker when young. The number of low-skilled workers is $\bar{L}$ and the number of high-skilled workers is $\bar{H}$. The efficiency units of labor provided by a high-skilled and a low-skilled agent are given by b_H and b_L, respectively. They are paid a wage per efficiency unit at rate, $w_l, l \in \{L, H\}$. Income $y^l = w_l b_l$ can be consumed in $t = 0$ or be saved and transformed to tomorrow's consumption possibilities. Agents are assumed to have quasi-homothetic preferences of the Stone-Geary form: Beyond a subsistence level to be expended they spend income on the good produced in the X-sector.[7] They have an instantaneous indirect utility function of the form $\log(e_t - \bar{e_t})$ where e_t is the expenditure for good X consumption and $\bar{e_t} \geq 0$ is the subsistence expenditure level in time t. Intertemporal preferences are assumed to be additive logarithmic with a discount factor δ.

The intertemporal problem consists of two parts: A saving decision and a portfolio choice. On the one hand, agents have to decide how much to expend on consumption, e_0, and how much to save, s. On the other hand, they have to put the saving in an appropriate portfolio of financial products. For this purpose they demand financial services. With the support of these services they decide how much of the saving is put into deposits, d, with a safe payoff r, and how much into risky state-contingent financial products (Arrow securities), f_θ, which pay off R_θ if state θ is realized and zero otherwise. We assume that all Arrow securities have the same expected payoff. Specifically, there exists $R > 0$ so that

$$R_\theta = \frac{R}{\pi_\theta}, \quad \theta \in \Theta. \tag{2.1}$$

[6] This structure is taken from Falkinger (2014).

[7] Achury et al. (2012) show that a Stone-Geary type utility function is appropriate for explaining stylized facts of household finance like higher saving rates of households with higher lifetime income or a larger fraction of risky assets in the portfolios of wealthy agents.

For transforming one unit of deposit, one unit of financial services from subsector 1 is needed; and for transforming one unit of Arrow securities, one unit of financial services from subsector 2 is required. Therefore, given the portfolio choice, $\{d, f\}$, with $f = \sum_{\theta\in\Theta} f_\theta$, agents have to pay a fee $T = p_{z_1} d + p_{z_2} f$ to the financial sector, where p_{z_1} and p_{z_2} are the prices for financial services Z_1 and Z_2, respectively.[8] Suppose the fee is charged in the first period and agents internalize the fee in their portfolio choice. The expected utility maximization problem of an agent l with income y^l is then given by:

$$\max_{s^l,\{f^l_\theta\}_{\theta\in\Theta},d^l} \mathbb{E}U = \log(e^l_0 - \bar{e}_0) + \delta \left[\mu \sum_{\theta\in\Theta} \pi_\theta \log(e^l_\theta - \bar{e}_1) + (1-\mu)\log(e^l_{\bar{\Theta}} - \bar{e}_1)\right]$$

s.t.

$$e^l_0 + (1+p_{z_1})d^l + (1+p_{z_2}) \sum_{\theta\in\Theta} f^l_\theta = y^l \tag{2.2}$$

$$e^l_\theta = \begin{cases} R_\theta f^l_\theta + r d^l, \text{ if } \theta \in \Theta \\ r d^l, \text{ otherwise} \end{cases} \tag{2.3}$$

$$s^l = \sum_{\theta\in\Theta} f^l_\theta + d^l. \tag{2.4}$$

In Section 2.5 aggregate demand functions for goods and financial services are derived from this program.

2.3.3 Production of goods (X-sector)

Firms in the X-sector employ low-skilled and high-skilled labor as input factors in a linear homogeneous production function

$$X = G^x(H_X, L_X),$$

where H_X, L_X denote respective labor employment in the X-sector. There is perfect competition with zero-profit prices. This means:

$$p_x = c_x(w_H, w_L), \tag{2.5}$$

where $c_x(w_H, w_L)$ are the unit costs and w_H, w_L are the wage rates per efficiency units.

[8] Without loss of generality, it was assumed that financial services are measured in units of savings. Without this normalization the cost of financial services per unit of saving would be $\tilde{p}_{z_i} = p_{z_i} n_i$ rather than p_{z_i}, where n_i denotes the units of financial services needed for one unit of saving in deposits ($i = 1$) and securities ($i = 2$), respectively.

The goods price is taken as numéraire, $p_x = 1$. Revenue X is distributed to labor as follows:

$$W_x = w_L L_x + w_H H_x = G^x(L_x, H_x),$$

where W_x is total wage earned in the X-sector.

Capital is used in technologies which transform savings into future consumption possibilities. Two types of technologies are available: A robust technology, which transforms under any condition (i.e., in Θ and $\bar{\Theta}$) one unit of capital invested today into r units of output tomorrow; furthermore, for $\theta \in \Theta$, a set of risky technologies specialized to θ-contingent environments. One unit of capital invested in technology θ delivers R_θ units of output if state $\theta \in \Theta$ occurs tomorrow and zero otherwise. Deposits are invested in the robust technology; savings in securities are invested in the respective risky technologies. The smaller the measure π_θ of the state to which a risky technology is targeted, the more productive the capital invested in the technology. Equation (2.1) expresses this relationship between specialization advantage and risk.

The separation of the production of old age consumption goods by capital from the labor based production of the goods consumed and invested in the active period of life is convenient from an analytical point of view. Under a more realistic perspective, however, capital is typically a prerequisite for producing with labor. In the extension in Section B.2.5, we show that essentially the same payoff structure arises if X is produced under monopolistic competition and capital is needed to set up firms – by robust and risky set-up technologies, respectively. Asset returns are then generated by the operating profits of the firms the set up of which has been financed by the asset.

In almost all of the further analysis only the relative payoff between robust and specialized risky technologies matters. It is given by:

$$\rho \equiv \frac{r}{R}.$$

The only exception is the discounting of future subsistence expenditure, $\frac{\bar{e_1}}{r}$, for which the level of the return on the robust technology matters.

2.3.4 Production of financial services (Z-sectors)

The financial sector Z consists of two subsectors, Z_1 and Z_2. They provide financial services for transforming savings through safe and risky assets into future consumption possibilities. (The assets are invested in the robust and risky technologies, and households get the generated revenue as return on their investment.) Z_i, $i \in \{1, 2\}$, is produced with

a linear homogeneous production function $G^{z_i}(.)$:

$$Z_i = G^{z_i}(H_{z_i}, L_{z_i}), \quad i \in \{1, 2\} \tag{2.6}$$

where H_{z_i}, L_{z_i} denote employment levels in the Z_i-sector.

In reality, fixed costs may play an important role in the provision of financial services. We consider such costs as an extension in Section B.2.1 and show how changes in fixed costs affect the equilibrium outcomes of our model.

We assume perfect competition in the Z-sectors and have therefore zero-profit prices

$$p_{z_i} = c_{z_i}(w_H, w_L), \quad i \in \{1, 2\} \tag{2.7}$$

where $c_{z_i}(w_H, w_L)$ are the unit costs.

Revenue $p_{z_i} Z_i,\ i \in \{1, 2\}$, is distributed to labor

$$W_{z_i} = w_L L_{z_i} + w_H H_{z_i} = p_{z_i} G^{z_i}(H_{z_i}, L_{z_i}), \quad i \in \{1, 2\}$$

where W_{z_i} is total labor income earned in the Z_i-sector.

As emphasized in the introduction, perfect competition in the Z-sector is an ideal benchmark rather than a description of reality. The role of rents is considered in the extension presented in Section B.2.2.

2.4 Production equilibrium and supply of goods and financial services

At the production side, the essential feature we want to address is variation in skill intensities. For an explicit comparative-static analysis we take production functions of the Cobb-Douglas form.

Let, for $j \in \{x, z_1, z_2\}$, G^j have Cobb-Douglas form

$$G^j\left(L_j, H_j\right) = A_j L_j^{1-\alpha_j} H_j^{\alpha_j},$$

where A_j is total factor productivity and α_j is the factor share of high-skilled workers in sector j.[9] Then

$$a_j^L = \frac{1}{A_j \kappa_j^{\alpha_j}}, \quad a_j^H = \frac{\kappa_j^{1-\alpha_j}}{A_j} \tag{2.8}$$

[9]The magnitudes of the total factor productivities depend on the unit in which financial services are measured. Since financial services are measured in units of savings, $A_x < A_{z_1} \leq A_{z_2}$ is a plausible restriction on total factor productivities. Analytically no such restriction is required for the results.

are the input coefficients, and cost-minimizing skill-intensities $\kappa_j \equiv a_j^H/a_j^L$ are given by

$$\kappa_j(\omega) = \frac{\gamma_j}{\omega}, \quad \gamma_j \equiv \frac{\alpha_j}{1-\alpha_j}, \tag{2.9}$$

where $\omega \equiv w_H/w_L$ is the relative wage per efficiency unit of skilled labor compared to unskilled labor, which reflects the skill premium (per efficiency unit).[10]

2.4.1 Wages and prices

We have for variable unit costs in sector j:

$$c_j\left(w_H, w_L\right) = \frac{w_L^{1-\alpha_j} w_H^{\alpha_j}}{A_j \Gamma_j}, \ \Gamma_j \equiv \alpha_j^{\alpha_j}(1-\alpha_j)^{1-\alpha_j}. \tag{2.10}$$

Using (2.10) and $p_x = 1$ in the zero-profit price equation (2.5), we obtain

$$w_L = A_x \Gamma_x \omega^{-\alpha_x}, \tag{2.11}$$

and from (2.7), for $i \in \{1, 2\}$,

$$p_{z_i} = \frac{A_x}{A_{z_i}} \frac{\Gamma_x}{\Gamma_{z_i}} \omega^{\alpha_{z_i} - \alpha_x}. \tag{2.12}$$

In sum, prices for financial services are related to the skill premium in the following way:

Fact 2.1. *The price of financial services Z_i, p_{z_i}, is an increasing function of ω if $\alpha_{z_i} > \alpha_x$. If $\alpha_{z_i} = \alpha_x$, then p_{z_i} is invariant with respect to ω. Moreover, $\alpha_{z_i} > \alpha_x$ ($\alpha_{z_i} = \alpha_x$) is equivalent to $\kappa_{z_i} > \kappa_x$ ($\kappa_{z_i} = \kappa_x$).*

As known from the Stolper-Samuelson theorem, this fact holds quite generally and is not an artifact of the Cobb-Douglas specification.

In the further analysis we make the following assumption about the factor intensity ranking of the three sectors.

Assumption 2.1. *$\alpha_{z_2} \geq \alpha_{z_1}$ and $\alpha_{z_1} \geq \alpha_x$ with at least one inequality holding strictly.*

In Section 2.8 we provide evidence on the sectoral skill intensities. Assumption 2.1 is consistent with the evidence.

[10] Note that $\kappa_j = \frac{b_H \bar{H}_j}{b_L \bar{L}_j}$. According to (2.9), the inverse labor demand function is $\omega = \left(\gamma_j \frac{b_L}{b_H}\right) \frac{\bar{L}_j}{\bar{H}_j}$. Thus, we have skill-biased technical change (in the sense of an outward shift of skilled-labor demand relative to unskilled-labor demand) if the output elasticity α_j of high-skilled labor rises or if there is low-skilled labor augmenting progress (that is b_L/b_H rises). It is worth noting that α_j is a sector-specific component whereas b_L/b_H is uniform across sectors.

2.4.2 Resource constraints

Total labor endowment in efficiency units is given by

$$L = b_L \bar{L}, \quad H = b_H \bar{H},$$

so that the "skill richness" of the total labor force is

$$k \equiv \frac{b_H \bar{H}}{b_L \bar{L}}.$$

The aggregate resource constraints are:

$$\begin{aligned} a_x^L X + a_{z_1}^L Z_1 + a_{z_2}^L Z_2 &= b_L \bar{L} \\ a_x^H X + a_{z_1}^H Z_1 + a_{z_2}^H Z_2 &= b_H \bar{H} \end{aligned} \tag{2.13}$$

with a_j^l, $j \in \{x, z_1, z_2\}$, $l \in \{H, L\}$ being functions of the skill premium ω defined in (2.9).

For illuminating the drivers of structural change on the production side it is worth looking, as an intermediary step, separately at the allocation of resources within the financial sector and the resource allocation between financial services and goods production. Let total employment (in efficiency units) in the financial sector be given by L_z and H_z, respectively, and denote by $k_z \equiv \frac{H_z}{L_z}$ the "skill richness" of the labor force in the financial sector. If $\alpha_{z_2} = \alpha_{z_1}$, the allocation of L_z and H_z on Z_1 and Z_2 is determined by the demand side only. If $\alpha_{z_2} > \alpha_{z_1}$, then we know from the Rybczynski analysis that both a rise in the skill premium ω and increased skill richness k_z shift resource allocation within the financial sector from Z_1 to Z_2. For the same reason, resources are shifted from goods production to the more skill-intensive provision of financial services if the skill premium or the skill richness rise in the economy.

In a general equilibrium, however, skill premium and employment in the financial sector are determined simultaneously with aggregate demand for financial services and goods.

2.5 Income distribution and aggregate demand

The demand for financial services comes from the need of agents to transform current savings into future income. For this purpose the asset-holding agents require financial products and expert services from the financial sector which support them by choosing and managing a portfolio of deposits and securities appropriate for their preferences.

The program max $\mathbb{E}U$ subject to (2.2)-(2.4) is only well-defined if $e_0 > \bar{e_0}$ and $e_1 > \bar{e_1}$.

This requires that

$$y^l = b_l w_l > \bar{y} \equiv \bar{e}_0 + (1 + p_{z_1}) \frac{\bar{e_1}}{r}, \quad l \in \{L, H\}. \tag{2.14}$$

$\bar{y}$ denotes the present value of future subsistence expenditure in units of today's final output.

Assuming $y^H \geq y^L$, which is equivalent to $\omega \geq b_L/b_H$, $y^L > \bar{y}$ is sufficient for (2.14). The following fact gives a necessary and sufficient condition for $y^L > \bar{y}$. The signs below the parameters show the sign of the respective partial derivatives.

Fact 2.2. *There exists a threshold* ω_L^+ *so that* $y^L > \bar{y}$ *if and only if* $\omega < \omega_L^+(\underset{+}{A_x}, \underset{+}{A_{z_1}}, \underset{+}{b_L}, \underset{-}{\bar{e}_0}, \underset{-}{\frac{\bar{e_1}}{r}})$.

Proof. Appendix B.1.3. □

Savings in securities is positive if and only if the following condition holds: $\mu R (1 + p_{z_1}) > (1 + p_{z_2}) r$. The condition can be rewritten in the form

$$\mu > p\rho, \quad p \equiv \frac{1 + p_{z_2}}{1 + p_{z_1}}, \quad \rho \equiv \frac{r}{R}. \tag{2.15}$$

$p\rho$ is the relative net payoff (i.e., after correction for costs of financial services) of savings in safe assets compared to savings in risky assets. If the condition is violated, the expected net payoff of risky investment is lower than the net payoff of risk-free investments and all saving is in deposits.

In the next subsection we analyze individual saving and expenditure behavior. Subsection 2.5.2 deals with aggregate demand.

2.5.1 Individual saving and expenditure behavior

As is derived in Appendix B.1.1, under the assumption that inequalities (2.14) and (2.15) are satisfied, individual savings in deposits and securities are given by

$$d^l = s_d \frac{\delta}{1 + \delta} \frac{y^l - \bar{y}}{1 + p_{z_1}} + \frac{\bar{e_1}}{r}, \quad l = \{L, H\}, \tag{2.16}$$

and

$$f^l = s_f \frac{\delta}{1 + \delta} \frac{y^l - \bar{y}}{1 + p_{z_2}}, \quad f_\theta^l = \pi_\theta f^l, \quad \theta \in \Theta, \quad l = \{L, H\}, \tag{2.17}$$

respectively, with

$$s_d = \frac{1 - \mu}{1 - p\rho}, \quad s_f = \frac{\mu - p\rho}{1 - p\rho}. \tag{2.18}$$

Apart from the savings for future subsistence expenditure, $\frac{\bar{e_1}}{r}$, in form of deposits, the saving level is proportional to the supernumerary budget $y^l - \bar{y}$. In real terms, the value

of the supernumerary budget, which is relevant as a basis for saving, depends on the price of the financial service charged on the particular form of savings – p_{z_1} for deposits and p_{z_2} for securities. The split of the saving on safe and risky assets is given by the marginal propensities to save in deposits, s_d, and in securities, s_f, respectively.[11] The propensity of safe investment increases in the relative net payoff of the safe asset, $p\rho$, and declines with the measure μ of states covered by securities. The propensity of risky investment reacts in the opposite direction.[12]

In contrast to net savings, gross savings include the fee to be paid for the financial services consumed in support for the transformation of savings into future income. Adding up $(1+p_{z_1})d^l + (1+p_{z_2})f^l$, we have

$$s^l + t^l = \frac{\delta}{1+\delta}(y^l - \bar{y}) + \frac{(1+p_{z_1})\bar{e_1}}{r}, \tag{2.19}$$

where $t^l = p_{z_1}d^l + p_{z_2}f^l$ denotes the total fee paid by agent l.

Current expenditures $e_0^l = y^l - (s^l + t^l)$ are thus:

$$e_0^l = \frac{1}{1+\delta}(y^l - \bar{y}) + \bar{e_0}. \tag{2.20}$$

For the discussion of structural change on the demand side, the effect of income on the portfolio structure is of particular importance.[13] According to (2.16) and (2.17), richer agents invest a larger share of their saving in risky assets than the relatively poorer ones. The reason is that the provision for future subsistence expenditure by safe investments has diminishing weight if people become richer. This means that saving in deposits has the character of a "necessity" and saving in risky securities is a "luxury". Moreover, if present subsistence expenditure is more pressing than future subsistence expenditure, people save a smaller part of their income when they are poor and the saving rate s/y rises when they get richer.[14] The following fact summarizes this important implication of our model.

---

[11] If inequality (2.15) is violated, then saving in securities is unattractive in the first place and we have a corner solution with $s_f = 0$ and $s_d = s = \frac{\delta}{1+\delta}\frac{y-\bar{y}}{1+p_{z_1}} + \frac{\bar{e_1}}{r}$.

[12] For $\bar{e_0} = \bar{e_1} = 0$ and $p_{z_1} = p_{z_2} = 0$, we have $s_d = \frac{1-\mu}{1-\rho}$ and $s_f = \frac{\mu-\rho}{1-\rho}$. Defining $\bar{R} = \frac{R}{\mu}$ and $\bar{\rho} = \frac{r}{\bar{R}}$, we can rewrite the two terms in the form $s_d = \frac{\bar{R}(1-\mu)}{\bar{R}-r/\mu}$ and $s_f = \frac{\mu\bar{R}-r/\mu}{\bar{R}-r/\mu}$. Thus, with Cobb-Douglas preferences and zero financial intermediation cost, the portfolio choice coincides with the one in Acemoglu and Zilibotti (1997), where the conditional expectation $\bar{R}$ of the productivity of risky technologies is used rather than the unconditional expectation R.

[13] Boppart (2015) analyzes the skill-content of the consumption basket of different income groups. With rising income, a household's demand shifts towards skill-intensive sectors (including financial services; also shown by Suellow (2015) in detail).

[14] The role of subsistence requirements for the saving behavior may call into mind the effects of fixed costs in the model of Greenwood and Jovanovic (1990), where saving rate and portfolio structure depend on an agent's wealth due to constrained participation in the use of financial intermediation service. While we consider the effect of a participation constraint as an extension in supplementary Section B.2, no such

Fact 2.3. *Let* $\bar{e}_0 > 0$ *or* $\bar{e}_1 > 0$.

a) If $\bar{e}_1 > 0$ *, then* $\frac{\partial(f/d)}{\partial y} > 0$.

b) For $\bar{e}_0 > 0$, $\frac{\partial(s/y)}{\partial y} > 0$ *if and only if* $\frac{\delta \bar{e}_0}{1+p_{z_1}} > \frac{\bar{e}_1}{r}\left[\frac{1+\delta}{s_d+s_f/p} - \delta\right]$. *(Note that for* $p = 1$ *the square-bracketed term reduces to one.)*

Proof. Part a) follows immediately from (2.16) and (2.17). For b) the definition of $\bar{y}$ in (2.14) is used. □

2.5.2 Aggregate demand for goods and financial services

Saving and expenditure behavior follow affine-linear functions. Therefore, aggregate behavior depends on two things: The level of aggregate income and the number of people over which the income is distributed. The latter comes in through the fact that subsistence requirements are bound to the existence of an agent, independent of her or his income.

Aggregating the two pools of agents, we have

$$N = \bar{L} + \bar{H}$$

for the size of the population and

$$W = w_L b_L \bar{L} + w_H b_H \bar{H}$$

for the level of aggregate income. In view of (2.11), the latter amounts to

$$W = A_x \Gamma_x b_L \bar{L} \omega^{-\alpha_x}(1 + \omega k). \tag{2.21}$$

The following fact shows that aggregate income, measured in units of X, is an increasing function of the skill premium ($\omega = w_H/w_L$).

Fact 2.4. *Under Assumption 2.1,* W *is increasing in* ω. *We have*

$$\frac{\partial W}{\partial \omega} = A_w \omega^{-\alpha_x}(1 - \alpha_x)(k - \kappa_x) > 0 \tag{2.22}$$

---

constraint exists in the baseline considered here. But everybody has to expend a certain sum to survive. This biases saving rate and portfolio structure. If people get richer the pressure of the subsistence requirements diminishes. There are of course other important differences to Greenwood and Jovanovic. In particular, all forms of saving require costly financial intermediation in our framework. Moreover, our focus is on inequality in labor income rather than wealth inequality and on structural change rather than growth.

with $A_w \equiv A_x \Gamma_x b_L \bar{L}$.

Proof. According to (2.21),

$$\begin{aligned}\frac{\partial W}{\partial \omega} &= A_w \omega^{-\alpha_x} \left[-\frac{\alpha_x}{\omega}(1+\omega k) + k \right] \\ &= A_w \omega^{-\alpha_x} \left[-\frac{\alpha_x}{\omega} + (1-\alpha_x)k \right] = A_w \omega^{-\alpha_x}(1-\alpha_x) \left[k - \frac{\alpha_x}{1-\alpha_x}\frac{w_L}{w_H} \right].\end{aligned}$$

According to (2.9),

$$\frac{\alpha_x}{1-\alpha_x} = \frac{w_H a_x^H}{w_L a_x^L}.$$

Thus, the square-bracketed term reduces to $k - \kappa_x$, which is positive if Assumption 2.1 holds. □

Financial services provision is more skill intensive than goods production, at least on average. Therefore, in terms of goods, aggregate wage income rises with the skill premium. A different matter is the impact of the skill premium on the purchasing power for financial services, the price of which rises too with the skill premium.

Aggregating individual investments in deposits, given by (2.16), we obtain

$$D = \left(s_d \frac{\delta}{1+\delta} \frac{\bar{w} - \bar{y}}{1 + p_{z_1}} + \frac{\bar{e_1}}{r} \right) N, \tag{2.23}$$

where $\bar{w} \equiv \frac{W}{N}$ denotes average income. In an analogous way, we have from (2.17):

$$F = s_f \frac{\delta}{1+\delta} \frac{\bar{w} - \bar{y}}{1 + p_{z_2}} N, \quad F_\theta = \pi_\theta F \tag{2.24}$$

for aggregate investments in securities and aggregate current expenditures are

$$E_0 = \left[\frac{1}{1+\delta}(\bar{w} - \bar{y}) + \bar{e_0} \right] N. \tag{2.25}$$

2.6 The effect of the skill premium on the sectoral structure

In a general equilibrium, sectoral structure and skill premium are determined simultaneously. As an intermediate step we characterize the sectoral structure as a function of the skill premium and exogenous parameters, keeping in mind that in the end the skill premium depends on exogenous parameters too. Not all possible values of skill premia and parameters are of interest, but only those which are reasonable candidates for a gen-

eral equilibrium, in which both financial sectors are viable, the subsistence of all agents is feasible and a positive skill premium results. The following paragraphs characterize the set of parameter configurations which guarantee these equilibrium properties.

Assumption 2.1 that financial service provision is more skill intensive than goods production ($\kappa_x < k < \kappa_z$) is equivalent to $\frac{\gamma_x}{k} < \omega < \frac{\gamma_z}{k}$ as we know from (2.9). At $\omega_{min} \equiv \frac{\gamma_x}{k}$ the Z-sector vanishes and beyond $\omega_{max} \equiv \frac{\gamma_z}{k}$ there would be no longer an X-sector. Hence, we consider the range $\omega \in (\omega_{min}, \omega_{max})$ in our search for the equilibrium skill premium.

Moreover, according to Fact 2.2, $\omega < \omega_L^+(\underset{+}{A_x}, \underset{+}{A_z}, \underset{+}{b_L}, \underset{-}{\bar{e_0}}, \underset{-}{\frac{\bar{e_1}}{r}})$ is required for guaranteeing subsistence for low-skilled agents. $\omega_L^+ \geq \omega_{max}$ holds if A_x, A_z and b_L are large enough (for given $\bar{e_0}, \frac{\bar{e_1}}{r}$), or $\bar{e_0}$ and $\frac{\bar{e_1}}{r}$ are not too high (for given A_x, A_z, b_L). If $\omega_L^+ < \omega_{max}$, only range $\omega \in (\omega_{min}, \omega_L^+)$ is feasible.

Finally, $\omega \geq b_L/b_H$ is required for $y^H \geq y^L$. This is guaranteed if $\omega_{min} \geq b_L/b_H$, which is equivalent to

$$\gamma_x \geq \frac{\bar{H}}{\bar{L}}.$$

In terms of exogenous fundamentals, the requirements mean that we restrict the possible combinations of exogenous model parameters

$$\xi = \left\{ A_x, A_{z_1}, A_{z_2}, \alpha_x, \alpha_{z_1}, \alpha_{z_2}, b_L, b_H, \bar{H}, \bar{L}, \bar{e_0}, \frac{\bar{e_1}}{r}, \rho, \mu, \delta \right\}$$

to the following set:

$$\Xi_0 \equiv \left\{ \xi \,\middle|\, \frac{\bar{H}}{\bar{L}} \leq \gamma_x, \frac{\gamma_x}{k} < \tilde{\omega}_{max} \right\}, \tag{2.26}$$

where $k = \frac{b_H \bar{H}}{b_L \bar{L}}$ and $\tilde{\omega}_{max} \equiv \min\left\{\omega_{max}, \omega_L^+(A_x, A_{z_1}, b_L, \bar{e_0}, \frac{\bar{e_1}}{r})\right\}$.

In general, the interaction of the allocation of resources between the X-sector and the Z-sector, on the one hand, and the allocation within the Z-sector on Z_1 and Z_2, on the other hand, are hard to disentangle in an economically transparent way. For qualitatively robust insights into important channels we have to reduce complexity on either the demand or the supply side. In the benchmark analysis presented in Section 2.6.1, 2.6.2 and 2.7, we shut down relative price effects within the financial sector by assuming identical technologies for Z_1 and Z_2.

Assumption 2.2. *$\alpha_{z_1} = \alpha_{z_2} = \alpha_z > \alpha_x$ and $A_{z_1} = A_{z_2} = A_z$.*[15]

[15]Without normalization $n_1 = n_2 = 1$, the assumption would read $\frac{A_{z1}}{n_1} = \frac{A_{z2}}{n_2}$. That is the provision of financial services per unit of saving must be equal in the two subsectors. For instance, new financial services may be provided more productively than traditional services, but, at the same time, more units of services are needed to transform a unit of saving into future payoff by complex rather than simple financial products.

Assumption 2.2 allows us to put focus on the income effects. In Appendix B.3 we consider the case $\alpha_{z_2} > \alpha_{z_1} = \alpha_x$ as a robustness check. Moreover, in the quantitative implementation of the model we solve the model numerically for α_j values that match U.S. data where $\alpha_{z_2} > \alpha_{z_1} > \alpha_x$.

We analyze first the impact of an increase in the skill premium on structural change within the financial sector.

2.6.1 Within change

The value added in subsector $Z_i, i = \{1, 2\}$, is equal to aggregate expenditure on the produced services. According to (2.23) and (2.24), aggregate expenditures for financial services have the following structure:

$$\frac{p_{z_2}F}{p_{z_1}D} = \frac{s_f\bar{\eta}(\omega)}{s_d\bar{\eta}(\omega) + \frac{1+\delta}{\delta}\frac{\bar{e_1}}{r}} \equiv \Phi(\underset{-}{s_d}, \underset{+}{s_f}, \frac{\bar{e_1}}{\underset{-}{r}}, \underset{+}{\bar{\eta}(\omega)}) \tag{2.27}$$

with $\bar{\eta}(\omega) \equiv \frac{\bar{w}(\omega)-\bar{y}}{1+p_z(\omega)}$. (Note that $p_{z_1} = p_{z_2} = p_z$ under Assumption 2.2.)

While the impacts of saving propensities s_d and s_f (defined in (2.18)) and the future subsistence requirements on the within structure are straightforward, the role of the skill premium is in general ambiguous. Apart from relative price effects, shut down by Assumption 2.2, the skill premium affects $\bar{\eta}(\omega)$ which is the average supernumerary income weighted by the cost of future subsistence.[16] It captures the income effect on within structural change. If $\bar{e_1} = 0$, there is no income effect on the demand structure for financial services. For $\bar{e_1} > 0$, however, the value-added ratio Φ of sector Z_2 compared to Z_1 depends on the skill premium in an U-shaped way. The following lemma characterizes the properties of $\bar{\eta}(\omega)$.

Lemma 2.1. *Let exogenous model parameters belong to* Ξ_0 *defined in* (2.26).

a) If $\xi \in \Xi_1 \equiv \Xi_0 \cap \{\xi | \alpha_x + \alpha_z > 1\}$, *then there exists a threshold* $\underset{?}{\underline{\omega}}(\underset{?}{A_x}, \underset{?}{A_z}, \underset{-}{\underline{k}}, \underset{?}{\frac{b_L\bar{L}}{N}}, \underset{-}{\bar{e_0}})$ *with* $\frac{\partial\bar{\eta}}{\partial\omega}\Big|_{\omega=\underline{\omega}} = 0$ *so that:*

$$\frac{\partial\bar{\eta}}{\partial\omega} < 0 \text{ for } \omega < \underline{\omega},$$
$$\frac{\partial\bar{\eta}}{\partial\omega} > 0 \text{ for } \omega > \underline{\omega}.$$

Especially, define $\Xi_D^1 \equiv \{\xi | \underline{\omega} > \omega_{min}\}$ *and* $\Xi_D^2 \equiv \{\xi | \underline{\omega} < \tilde{\omega}_{max}\}$. *If* $\xi \in \Xi_1 - \Xi_D^1$, *then* $\frac{\partial\bar{\eta}}{\partial\omega} > 0$ *for all* $\omega \in (\omega_{min}, \tilde{\omega}_{max})$. *If* $\xi \in \Xi_1 - \Xi_D^2$, *then* $\frac{\partial\bar{\eta}}{\partial\omega} < 0$ *for all* $\omega \in (\omega_{min}, \tilde{\omega}_{max})$.

[16] See discussion in Section B.3 for the case of changing relative prices within the Z-sector.

b) For the comparative static analysis we have:

$$\bar{\eta}\left(\omega \,\middle|\, \underset{+}{A_x}, \underset{+}{A_z}, \underset{+}{k}, \underset{+}{\frac{b_L\bar{L}}{N}}, \underset{-}{\bar{e_0}}, \underset{-}{\frac{\bar{e_1}}{r}}\right)$$

Proof. Appendix B.1.3. □

On the one hand, a higher ω raises the average wage. On the other hand, the prices of financial services are increasing, which has a negative effect on the purchasing power. According to Lemma 2.1, the first effect dominates if the skill premium is sufficiently high.

In sum, we have the following partial results about within structural change in the finance sector.

Proposition 2.1. *Let* $\bar{e_1} > 0$.

a) A rise in the skill premium leads to structural change from subsector Z_1 *to subsector* Z_2 *(in terms of value-added) at high levels of the skill premium* ($\omega > \underline{\omega}$) *and to structural change from* Z_2 *to* Z_1 *at low levels of skill premium.*

b) For a given skill premium, a rise of $A_x, A_z, k, \frac{b_L\bar{L}}{N}$ *or a decline of* $\bar{e_0}, \frac{\bar{e_1}}{r}$ *lead to structural change from* Z_1 *to* Z_2*. A rise of* μ *or a decline of* ρ *also lead to change from* Z_1 *to* Z_2*, even if* $\bar{e_1} = 0$.

Proof. (2.27), Lemma 2.1 and the fact that a rise in μ or a decline in ρ raise s_f (at cost of s_d). □

The proposition describes only a partial effect. For a full comparative-static equilibrium analysis, we have to combine the direct effects of exogenous fundamentals with their indirect effects through the equilibrium skill premium. We come back on the total effects in Section 2.7.4.

2.6.2 Between change

For $\alpha_{z_1} = \alpha_{z_2} = \alpha_z$ and $A_{z_1} = A_{z_2} = A_z$, aggregate supply of financial services reduces to:

$$Z(= Z_1 + Z_2) = A_z L_z \kappa_z^{\alpha}.$$

The allocation between the X-and the Z-sector is then determined by the resource constraints:

$$a_x^L X + a_z^L Z = b_L \bar{L},$$
$$a_x^H X + a_z^H Z = b_H \bar{H}.$$

Solving the system for X and Z, we obtain

$$X = \frac{b_L \bar{L}}{a_x^L} \frac{\kappa_z - k}{\kappa_z - \kappa_x}, \quad Z = \frac{b_L \bar{L}}{a_z^L} \frac{k - \kappa_x}{\kappa_z - \kappa_x}. \tag{2.28}$$

As a result, we have for values of the services supplied by the financial sector compared to the output of the goods sector:

$$\frac{p_z Z}{X} = \frac{p_z(\omega) a_x^L(\omega)}{a_z^L(\omega)} \frac{k - \kappa_x(\omega)}{\kappa_z(\omega) - k} \equiv \underset{+}{\Psi}(\underset{+}{\omega}, \underset{+}{k}). \tag{2.29}$$

This gives us the following result for the comparative-static effects on the supply structure. [17]

Proposition 2.2. *An increase in the skill premium shifts the supply structure from goods production to financial services provision. An increase in the high skilled labor share* (k) *has the same effect.*

Proof. The signs of the respective partial derivatives in (2.29) follow from $\kappa_z > \kappa_x$, the Rybczynski analysis and the fact that p_z rises in ω. □

The proposition characterizes the supply structure as a function of exogenous fundamentals and the skill premium. The supply structure interacts with demand, which depends on aggregate income and prices and thus also reacts to the skill premium. To close the analysis, we have to determine the equilibrium skill premium. Section 2.7.3 will then summarize the general equilibrium effect of the skill premium on the between sectoral structure.

[17] Note that (2.29) characterizes the supply structure of labor produced output. If capital is used as set-up capital as in the extended model in Section B.2.5, then X is indeed the total size of final output in the goods sector. In the baseline model considered here there is in addition the output generated for old age consumption by past capital investments. Thus, the total size of goods transactions becomes $\bar{X} \equiv X + rD + \mu RF$ with $X = E_0 + S$, $S = D + F$ and the between structural change ratio is $\bar{\Psi} \equiv \frac{p_z D + p_z F}{\bar{X}} = \frac{p_z D + p_z F}{X + rD + \mu RF}$ with D, F and E_0 from (2.23)-(2.25). It is, ceteris paribus, increasing in ω if $S'E_0 - SE_0' - (\mu R - r)(DF' - FD') > 0$ where D', F', S' and E_0' are the respective derivatives with respect to ω. This means, if the between change $(S'E_0 - SE_0')$ is larger than within change $(DF' - FD')$ multiplied with the return difference $(\mu R - r)$.

2.7 General equilibrium

Aggregate demand in the X-sector is composed of consumer goods demand, E_0, and investment goods demand, $S = D + F$. On top of it, old agents consume the output generated by the capital they invested in the period before.

Aggregating the individual budget constraints (2.2), we obtain:

$$E_0 + D + F + p_{z_1} D + p_{z_2} F = W, \tag{2.30}$$

where $W = W_x + W_z$, $W_x = X$ and $W_z = p_{z_1} G^{z_1}(H_{z_1}, L_{z_1}) + p_{z_2} G^{z_2}(H_{z_2}, L_{z_2})$. If the Z_1 and Z_2-markets are cleared, we have $G^{z_1}(H_{z_1}, L_{z_1}) = D$ and $G^{z_2}(H_{z_2}, L_{z_2}) = F$ so that (2.30) reduces to

$$E_0 + D + F = X.$$

Thus, the goods market is automatically cleared if the markets for financial services are cleared.

Aggregate demand for financial services comes from savings in deposits D and savings in securities F. Adding up (2.23) and (2.24), we have for aggregate demand in the Z-sector

$$Z^D = \left(\frac{\delta}{1+\delta} \frac{\bar{w} - \bar{y}}{1 + p_z} + \frac{\bar{e_1}}{r} \right) N. \tag{2.31}$$

From (2.28) we know that aggregate Z-supply in a production equilibrium is

$$Z^S = A_z b_L \bar{L} \kappa_z^{\alpha_z} \frac{k - \kappa_x}{\kappa_z - \kappa_x} \tag{2.32}$$

where $a_z^L = \frac{1}{A_z \kappa_z^{\alpha_z}}$ was used.

2.7.1 Existence, uniqueness and stability of equilibrium

Both market sides are functions of ω (which works through $\bar{w}$ and p_z on the demand side and through skill intensities κ_x, κ_z on the supply side). For a stable equilibrium, the condition

$$\frac{dZ^D}{d\omega} < \frac{dZ^S}{d\omega} \tag{2.33}$$

is required at the market clearing ω-value. (Since p_z is increasing in ω, inequality (2.33) guarantees that a rise in price p_z goes hand in hand with a reduction of excess demand and a fall in the price reduces excess supply.)

The supply function is characterized by the following fact.

Fact 2.5. *Z^S is an increasing strictly concave function of ω starting at $\lim_{\omega\to\omega_{min}} Z^S = 0$ and approaching $A_z b_L \bar{L} k^{\alpha_z}$ at ω_{max}. More specifically,*

$$Z^S = A_z b_L \bar{L} \frac{\gamma_z^{\alpha_z}}{\gamma_z - \gamma_x} g(\underset{+}{\omega}, \underset{+}{k}), \quad g(\omega, k) \equiv \omega^{-\alpha_z}(k\omega - \gamma_x). \tag{2.34}$$

Proof. Appendix B.1.3. □

For the demand side the following fact applies.

Fact 2.6. *Aggregate demand for financial services is given by:*

$$Z^D = \left[\frac{\delta}{1+\delta} \bar{\eta} \left(\omega \left| \underset{+}{A_x}, \underset{+}{A_z}, \underset{+}{k}, \underset{+}{\frac{b_L \bar{L}}{N}}, \underset{-}{\bar{e_0}}, \underset{-}{\frac{\bar{e_1}}{r}} \right. \right) + \frac{\bar{e_1}}{r} \right] N,$$

where $\bar{\eta}$ was discussed in Lemma 2.1. For all $\xi \in \Xi_1$, Z^D is defined and positive on the ω-domain $(\omega_{min}, \tilde{\omega}_{max})$. Moreover, it is either U-shaped in ω (for $\xi \in \Xi_D \equiv \Xi_1 \cap \Xi_D^1 \cap \Xi_D^2$), increasing over the entire domain (for $\xi \in \Xi_1 - \Xi_D^1$) or declining for all ω (if $\xi \in \Xi_1 - \Xi_D^2$).

Proof. Equation (2.31) and Lemma 2.1. □

Figure 2.4 shows in the (ω, Z)-space the supply and demand curves under the assumption that

$$Z^D(\tilde{\omega}_{max}) < Z^S(\tilde{\omega}_{max}), \tag{2.35}$$

where $\tilde{\omega}_{max}$ was defined in (2.26).[18]

If inequality (2.35) holds, then the market clearing condition $Z^D(\omega) = Z^S(\omega)$ has a unique solution ω^* within $(\omega_{min}, \tilde{\omega}_{max})$. Moreover, stability condition (2.33) is fulfilled at ω^*. This establishes the following proposition.

Proposition 2.3. *Define $\Xi_E = \Xi_1 \cap \left\{ \xi | Z^D(\tilde{\omega}_{max}) < Z^S(\tilde{\omega}_{max}) \right\}$. For all $\xi \in \Xi_E$, there exists a unique and stable equilibrium.*

Proof. Continuity of Z^D on $\omega \in (\omega_{min}, \tilde{\omega}_{max})$ and properties of the shape of Z^D established in Fact 2.6. □

2.7.2 Equilibrium skill premium

For the comparative-static equilibrium analysis, we have to look at the excess demand function $Z^D - Z^S$. Because of stability condition $\frac{\partial(Z^D - Z^S)}{\partial \omega} < 0$, we know that for any

[18] If $\tilde{\omega}_{max} = \omega_L^+$, then $Z^D(\tilde{\omega})$ is to be read as $Z^D(\omega) < Z^S(\omega)$ for all $\omega < \omega_L^+ - \epsilon$, with ϵ arbitrarily small. Figure 2.4 assumes $\xi \in \Xi_D$; yet, from Fact 2.6 it is obvious that for $\xi \in \Xi_1 - \Xi_D^1$ the Z^D-curve would cross the Z^S-curve at ω^* as in Case I, whereas for $\xi \in \Xi_1 - \Xi_D^2$ we would have at ω^* the situation illustrated in Case II.

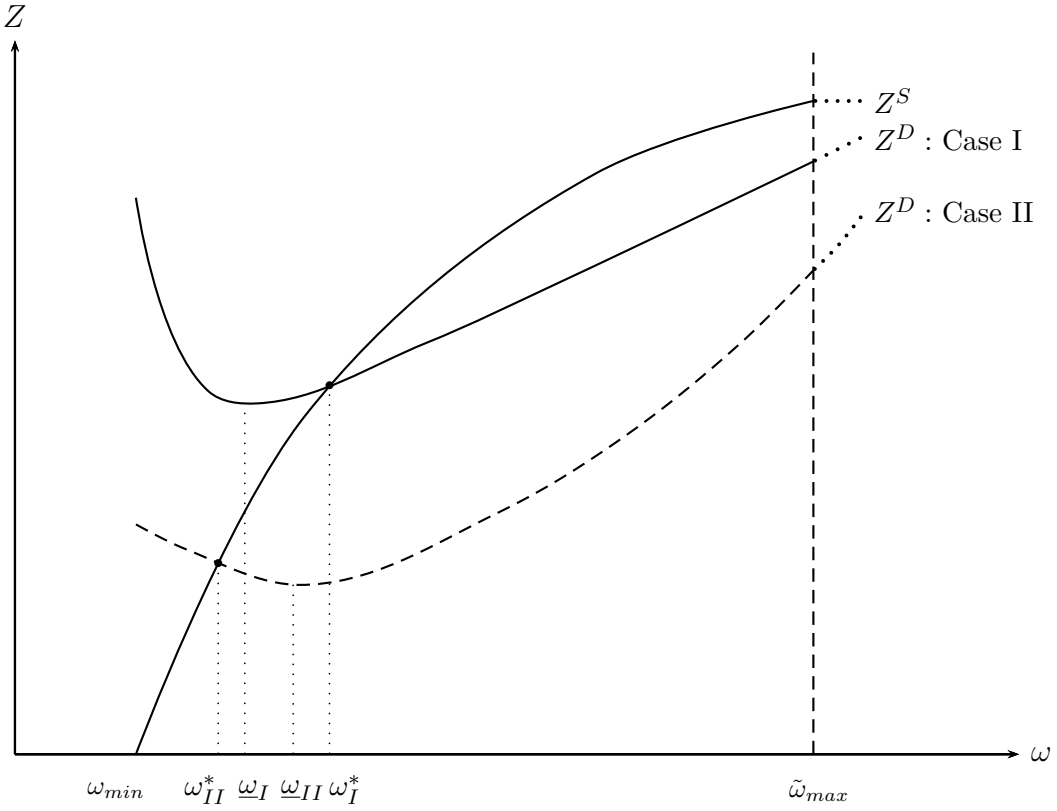

Figure 2.4: Equilibrium in the financial service sector.

exogenous change of a component $\mathfrak{i}$ of $\xi \in \Xi_E$

$$sign\frac{\partial\omega^*}{\partial\mathfrak{i}} = sign\frac{\partial(Z^D - Z^S)}{\partial\mathfrak{i}}\,|_{Z^D=Z^S}.$$

For signing the impact of exogenous fundamentals on the equilibrium, we express excessive demand explicitly as a function of model parameters. Using (2.21) and (2.12), we have

$$\frac{\bar{w}N}{1+p_z} = A_x b_L \bar{L} D_1(\omega|\underset{+}{\frac{A_z}{A_x}}, \underset{+}{k}), \tag{2.36}$$

where $D_1 \equiv \frac{\Gamma_x(1+\omega k)}{\omega^{\alpha_x} + \frac{A_x\Gamma_x}{A_z\Gamma_z}\omega^{\alpha_z}}$ and the signs below parameters in (2.36) express the signs of their impact on D_1. Term D_1 captures the purchasing power effect.

Moreover, substituting (2.12) for p_{z_1} in (2.14) we can write the term $\frac{\delta}{1+\delta}\frac{\bar{y}}{1+p_z} - \frac{\bar{e_1}}{r}$ in the form:

$$D_0(\omega|\underset{+}{\frac{A_z}{A_x}}, \underset{+}{\bar{e_0}}, \underset{-}{\frac{\bar{e_1}}{r}}) = \frac{1}{1+\delta}\left[\frac{\delta\bar{e_0}}{1+\frac{A_x\Gamma_x}{A_z\Gamma_z}\omega^{\alpha_z-\alpha_x}} - \frac{\bar{e_1}}{r}\right]. \tag{2.37}$$

Term D_0 captures the effect of the subsistence requirements on the aggregate demand

for financial services. The sign of the square-bracketed term is positive if the present subsistence expenditure $\bar{e_0}$ dominates the future subsistence expenditure $\bar{e_1}$. It is negative if $\bar{e_1}$ dominates $\bar{e_0}$. For the economic interpretation of the relevant notion of dominance it is useful to recall $\frac{A_x\Gamma_x}{A_z\Gamma_z}\omega^{\alpha_z-\alpha_x} = p_z$. Thus $D_0(\omega|\frac{A_z}{A_x}, \bar{e_0}, \frac{\bar{e_1}}{r}) > 0$ $(=, < 0)$ if and only if

$$\frac{\delta\bar{e_0}}{1+p_z} > \frac{\bar{e_1}}{r} \ (=, < \frac{\bar{e_1}}{r}, \text{ resp.}). \tag{2.38}$$

This is exactly the condition for a rising (constant, declining, resp.) saving rate derived in Fact 2.3.b). (Note that $p = 1$ in the benchmark case.) If present subsistence expenditures are more pressing than future ones, people save more and demand more financial services if they become richer and get farther away from subsistence problems.

Using D_0 and (2.36) in (2.31) and combining the result with (2.34), we conclude that $Z^D - Z^S$ is equal to the term

$$A_x b_L \bar{L}\left[\frac{\delta}{1+\delta}D_1(\omega|\underset{+}{\frac{A_z}{A_x}}, \underset{+}{k}) - \frac{N}{A_x b_L \bar{L}}D_0(\omega|\underset{+}{\frac{A_z}{A_x}}, \underset{+}{\bar{e_0}}, \underset{-}{\frac{\bar{e_1}}{r}}) - \frac{A_z}{A_x}\frac{\gamma_z^{\alpha_z}}{\gamma_z-\gamma_x}\underset{+}{g(\omega, k)}\right]. \tag{2.39}$$

Hence, $\bar{e_1}$ has a positive impact on $Z^D - Z^S$ and thus on ω^*; $\bar{e_0}$ has a negative impact. $\frac{A_z}{A_x}$ and k have opposing effects so that their impacts cannot be signed unambiguously by inspection of (2.39).

The most interesting question is how technical change affects the equilibrium skill premium. For this we have to look at the impact of $\frac{A_x b_L \bar{L}}{N}$ on $Z^D - Z^S$. (Since $\frac{A_z}{A_x}$ has an ambiguous effect, we only consider uniform progress across sectors, that is, total factor productivity A_z rises pari passu with A_x.) The answer depends on condition (2.38). If $\frac{\delta\bar{e_0}}{1+p_z} > \frac{\bar{e_1}}{r}$, D_0 is positive and ω^* increases if $\frac{A_x b_L \bar{L}}{N}$ rises. If $\frac{\delta\bar{e_0}}{1+p_z} < \frac{\bar{e_1}}{r}$, then D_0 is negative and ω^* declines if $\frac{A_x b_L \bar{L}}{N}$ increases. For $\bar{e}_0 = \bar{e}_1 = 0$, $\frac{A_x b_L \bar{L}}{N}$ has no effect.

In sum, we have the following partial effects of the parameters on the equilibrium skill premium:[19]

$$\omega^*(\underset{?}{\frac{A_z}{A_x}}, \underset{?}{k}, \underset{+/-}{\frac{A_x b_L \bar{L}}{N}}, \underset{-}{\bar{e_0}}, \underset{+}{\frac{\bar{e_1}}{r}}), \tag{2.40}$$

where the impact of $\frac{A_x b_L \bar{L}}{N}$ depends on the cases discussed above.

All addressed effects refer to the partial derivatives, that is, they hold under the condition that other parameters do not change simultaneously. Economically this means, the effects come from a single source. In particular, for the effect of $\frac{b_L \bar{L}}{N}$ on ω^*, skill richness

[19]The signs below the parameters represent the partial derivatives. The combination $+/-$ is used for pointing to case-dependent impacts. A question mark means that the impact of the respective parameter cannot be signed without further investigation.

$k = \frac{b_H \bar{H}}{b_L \bar{L}}$ is held constant in the comparison. This requires a careful interpretation of the described effect of $\frac{b_L \bar{L}}{N}$. The following fact provides an economically meaningful description of the variations which are consistent with a constant k and a rise in $\frac{b_L \bar{L}}{N}$.

Fact 2.7. *A rise in $\frac{b_L \bar{L}}{N}$ is consistent with a constant k if there is:*

a) Uniform factor-augmenting technical progress, raising b_L pari passu with b_H.

b) A shift in labor supply from unskilled to skilled labor accompanied by factor augmenting progress that is biased towards the low-skilled. (Note that such low-skilled labor augmentation depresses the relative wage of the unskilled – like skill-biased technical change.)

Proof. Use $N = \bar{L} + \bar{H}$ for $\frac{N}{b_L \bar{L}} = \frac{1 + \frac{\bar{H}}{\bar{L}}}{b_L}$. Hence, $k = \frac{b_H \bar{H}}{b_L \bar{L}}$ remains constant under a decrease in $\frac{N}{b_L \bar{L}}$ if either b_L and b_H rise proportionally and $\bar{H}/\bar{L}$ does not change or $\frac{\bar{H}}{\bar{L}}$ rises and b_L rises such that $\frac{b_L}{b_H}$ grows proportionally to $\frac{\bar{H}}{\bar{L}}$. □

With these clarification the following proposition summarizes the comparative static equilibrium results.

Proposition 2.4. *Let $\bar{e}_0 >$ or $\bar{e}_1 > 0$.*

a) Uniform productivity growth across sectors (raising A_x and A_z proportionally) or uniform factor-augmenting technical progress (raising b_L and b_H proportionally) have a positive effect on the equilibrium skill premium if the present subsistence expenditure dominates the future subsistence expenditure; if the future subsistence expenditure dominates, then the skill premium declines.

b) A shift of labor supply from unskilled to skilled work accompanied by factor augmentation which is biased towards low-skilled labor has the same effect on the equilibrium skill premium as factor augmenting progress that is uniform.

c) The equilibrium skill premium rises, if future subsistence expenditure ($\bar{e}_1$) increases or present subsistence expenditure ($\bar{e}_0$) declines.

Proof. Fact 2.7 and main text. □

For the economic intuition behind a) and b) it is useful to remember Fact 2.3.b). If present subsistence expenditure weighs more than future subsistence requirements then the saving rate and therefore demand for financial services are rising with income. Since the financial services are more skill intensive than goods, this rise of demand induces a rise in the skill premium. The rising income in turn comes from technical progress or a better educated workforce. The intuition for c) is: If future subsistence expenditure is

high, agents have to save more and need more financial services; and if present subsistence expenditure is low, they can afford to save more and to spend more for financial services.

It is worth noting that positive subsistence expenditure ($\bar{e}_0 > 0$ or $\bar{e}_1 > 0$) is essential for the comparative-static results stated in Proposition 2.4. For $\bar{e}_0 = \bar{e}_1 = 0$, expression (2.39) boils down to

$$A_x b_L \bar{L} \left[\frac{\delta}{1+\delta} D_1 \left(\omega \left| \underset{+}{\frac{A_z}{A_x}}, \underset{+}{k} \right. \right) - \frac{A_z}{A_x} \frac{\gamma_z^{\alpha_z}}{\gamma_z - \gamma_x} g \left(\omega, \underset{+}{k} \right) \right].$$

Thus, uniform productivity growth has no effect in this case nor has $\frac{b_L \bar{L}}{N}$.

2.7.3 Structural change between production and financial service sectors

Combining the results of subsections 2.7.2 and 2.6.2, we obtain the following results for the structural change between production and financial services in equilibrium:

Proposition 2.5. *For all $\xi \in \Xi_E$, at given $\frac{A_z}{A_x}$, k, any change in other exogenous fundamental which raises (lowers) the skill premium leads to structural change from X to Z (Z to X, respectively).*

Proof. Equation (2.29). Since p_z rises with ω, the rise of ψ immediately implies that $\frac{p_z Z}{X}$ rises too. □

2.7.4 Structural change within the financial sector

Finally, for structural change within the financial sector, we have the following results in equilibrium:

Proposition 2.6. *Let $\underline{\omega}$ be the threshold defined in Lemma 2.1 and parameters fulfill $\xi \in \Xi_E$. Then, under the assumption that prices do not differ across financial services, the following comparative static results hold for structural change within the financial sector as long as $\bar{e}_1 > 0$:*

a) At high levels of the skill premium ($\omega^* > \underline{\omega}$), a fall of $\bar{e}_0$ leads to a shift from Z_1 to Z_2. In addition, if present subsistence expenditure dominates future subsistence expenditure, uniform productivity growth across sectors (i.e. a proportional rise of A_x and A_z) as well as an increase in $\frac{b_L \bar{L}}{N}$ change the structure within the financial sector from Z_1 towards Z_2. According to Proposition 2.4 and 2.5, these changes induce an increase in the inequality level ω^*, accompanied by a simultaneous structural change from the goods to the financial service sector.

b) At low levels of the skill premium ($\omega^* < \underline{\omega}$), a fall of $\bar{e_1}$ leads to a shift from Z_1 to Z_2. In addition, if future subsistence expenditure dominates present subsistence expenditure, uniform productivity growth across sectors as well as and an increase in $\frac{b_L \bar{L}}{N}$ change the structure within the financial sector from Z_1 towards Z_2. However, according to Proposition 2.4 and 2.5, these changes correspond to a decrease in the inequality level ω^*, accompanied by a simultaneous a structural change from the financial service to the goods sector.

c) Financial product innovation (a rise of μ) or rising attractiveness of risky investments (a decline of ρ) lead to structural change from Z_1 to Z_2, even if $\bar{e}_1 = 0$.

Proof. Using (2.27), (2.40), and Lemma 2.1, we have

$$\Phi \left\{ \underset{-}{s_d}, \underset{+}{s_f}, \underset{-}{\frac{\bar{e_1}}{r}}, \underset{+}{\bar{\eta}} \left[\underset{+/-}{\omega^*} \left(\underset{?}{\frac{A_z}{A_x}}, \underset{?}{k}, \underset{+/-}{\frac{A_x b_L \bar{L}}{N}}, \underset{-}{\bar{e_0}}, \underset{+}{\frac{\bar{e_1}}{r}} \right), \underset{+}{A_x}, \underset{+}{A_z}, \underset{+}{k}, \underset{+}{\frac{b_L \bar{L}}{N}}, \underset{-}{\bar{e_0}}, \underset{-}{\frac{\bar{e_1}}{r}} \right] \right\},$$

where the signs below the parameters show the sign of the respective partial derivative of the functions $\Phi\{\cdot\}$, $\bar{\eta}[\cdot]$ and $\omega^*(\cdot)$. The plus below ω^* applies for $\omega^* > \underline{\omega}$, the minus for $\omega^* < \underline{\omega}$. The plus below $\frac{A_x b_L \bar{L}}{N}$ applies for the case that $\bar{e}_0$ dominates $\bar{e}_1$; the minus applies if $\bar{e}_1$ dominates $\bar{e}_0$. For the impacts of μ and ρ note that s_f is rising and s_d is declining in μ and rising in ρ. □

It is worth noting that for $\bar{e_1} = 0$ there is no income effect on the portfolio structure so that the channel between skill premium and financial structure is shut down. Since in the benchmark considered here relative price effects within the financial sector were shut down too, for $\bar{e_1} = 0$ only financial innovation (a rise in μ) and rising relative returns on risky investment (a decline of ρ) remain as sources of structural change within the financial sector. This changes in the model variant with different technologies for Z_1 and Z_2 considered in Appendix B.3.

The punchline of the general equilibrium analysis in the baseline model is: When the skill premium has reached a certain level, a rise in average income leads to rising inequality and to twofold structural change towards and within the financial sector simultaneously. The rise in income can be triggered by a general rise of productivity or by an increased selection of the population into higher education (accompanied by labor augmenting progress that makes low-skilled labor abundant relative to skilled labor). The income effects generated by technical progress or education are robust drivers of the developments outlined at the beginning of this paper. They can explain a rising skill premium and the twofold structural change towards and within finance by a single source, holding everything else constant. Yet, of course, in reality the effects triggered by this source are

overlaid by many other things that happen at the same time. The model points to a series of other exogenous fundamentals that affect skill premium and economic structure. Thus, the specific combination of determinants that actually determine the observed patterns of inequality and structural change can only be identified by empirical analysis. The quantitative analysis in Section 2.8 illustrates possible combinations of exogenous factors which are consistent with the development observed from 1980 onwards.

2.7.5 Distortions

The main arguments why financial markets are distorted by imperfections are: First, the complexity of new financial products confuses people. Second, the expertise required to deal with the complex products gives to the agents within the financial sector an advantage that can be exploited for extracting rents from their clients. Third, not all households will be able to participate in the security markets. How would such imperfections change the equilibrium outcome qualitatively?

If confusion leads to wrong beliefs about the opportunities provided by securities, the consequences are straightforward. For instance, if investors are euphoric about the measure μ of risky states covered by state-contingent financial products, then, according to (2.18), the propensity s_f for new financial services rises while the propensity to save in deposits declines. As a consequence, the new finance sector gains weight compared to traditional financial services, as shown by (2.27). This structural change within the financial sector does not affect total demand Z^D (see (2.31)) so that equilibrium skill premium and financial sector share do not change compared to the benchmark analysis. Misperception of the relative returns on risky investments ρ would affect the equilibrium outcome in a similar way. In sum, euphoric beliefs about measure or performance of state-contingent financial products enhance structural change within the financial sector.

The effect of rents in the financial sector are in general more complex. If they are extracted by charging to clients a mark up on the costs c_z of providing financial services, relative prices are distorted so that all equilibrium values are affected (see supplementary Section B.2 for more details). Such distortive allocative effects are excluded if rents are earned by charging to clients a lump sum fee τ for providing financial services on top of the price p_z for covering costs c_z of providing the services. In this case, an unambiguous redistributive effect raises the finance share in total income. The effect on the structure within the financial sector depends on the way in which the rents are distributed on the financial agents.

Fixed fees have more far-reaching consequences if they exclude low-income earners from participating in a financial market. In the presented model, low-skilled agents would be excluded from using new financial services if a lump sum fee τ is charged to clients

of the Z_2-sector which is higher than the supernumerary income $y^L - \bar{y}$. Under such a participation constraint, low-skilled workers invest all their savings in deposits. In contrast, high-skilled workers can pay fee τ for participating in the Z_2-market. If the lump sum fee τ corresponds to a real fixed cost arising in the provision of financial services, then τ has similar effects like subsistence expenditure $\bar{e_0}$. If however τ is charged to generate rents for the high-skilled workers in the financial sector, we have redistribution of income among high-skilled workers. Since for Stone-Geary preferences such redistribution does not change aggregate savings, equilibrium wage premium ω and finance ratio Ψ remain unchanged compared to the benchmark analysis. Yet, due to the participation constraint the ratio of new finance compared to traditional finance changes to

$$\tilde{\Phi} = \frac{s_f \beta_H}{1 - s_f \beta_H + \frac{1+\delta}{\delta} \frac{1+p_z}{\bar{w}-\bar{y}} \frac{\bar{e_1}}{r}} \tag{2.41}$$

where $\beta_H \equiv \frac{y^H - \bar{y}}{\bar{w}-\bar{y}} \frac{\bar{H}}{N} < 1$ is the income share of high-skilled agents.[20] Comparing (2.41) with (2.27), we see two things: First, since only part of the population participates in the Z_2-market the new finance share is lower than under full participation. Second and more interesting, the distribution of income becomes an important determinant of the structure within the financial sector. If the income share (β_H) of high-skilled agent rises then the new finance share rises, too.

2.8 Empirical evidence and numerical exercises

In this section we first provide empirical evidence on the twofold structural change and on wage inequality and then we carry out numerical exercises to show how our model can replicate the observed changes.

2.8.1 Empirics

2.8.1.1 Data

We use data from the Current Population Survey (March CPS) for the survey years 1980-2013 from IPUMS-CPS by King et al. (2010).[21] This data set allows us to split the sampled population (weighted with the sampling weight) into our three sectors and two skill levels: The X-sector consists of all sectors of the U.S. economy except finance. The finance

[20] See Appendix B.2 for details.

[21] Survey years 1980-2013 represent years 1979-2012 because households are surveyed about last year's job. This means whenever we talk about a year the data considered represent the situation a year before.

sector is finance and insurance without real estate.[22] "Traditional finance" Z_1 includes banking, credit agencies and insurance. "New finance" Z_2 is security and commodity brokerage and investment companies. We define a worker (who worked positive weeks last year) to be high-skilled if she/he holds a college degree (four-year college) or more. Then, $\bar{H}_j$ is the number of high-skilled workers in sector $j \in \{x, z_1, z_2\}$ and $\bar{L}_j$ is the number of low-skilled workers in sector $j \in \{x, z_1, z_2\}$. For each skill level, we calculate for the three sectors the average yearly hours worked last year (i.e., h_j^l, $j \in \{x, z_1, z_2\}$, $l \in \{H, L\}$) and the respective average hourly real wages (i.e., w_j^l, $j \in \{x, z_1, z_2\}$, $l \in \{H, L\}$).[23]

In our data analysis we use "actual" and "normalized" numbers for employment and wage levels. The "actual" numbers use the observed sector- and skill-specific average yearly hours worked and the respective average hourly wage. The "normalized" numbers are calculated all with the same basis of hours worked and hourly wage (i.e., the ones from the X-sector).[24] The normalization allows us to separate the effects we can identify in the theoretical, frictionless model from two frictions observed in reality: (i) Low- and high-skilled Z-workers work more hours per year than low- and high-skilled X-workers. More precisely, for the U.S. over the last decades on average a Z-worker has worked about 9% more than a X-worker. (ii) There is the finance premium on hourly wages for low- and high-skilled Z-workers.[25] CPS data show that the finance premium increased over time and differs for the two subsectors: In Z_1 workers earn about 15% more than in the X-sector, in Z_2 it is even 50%.

The sectoral structure-figures below show black and gray lines: The gray lines corre-

[22]This corresponds to the standard classification as in Philippon and Reshef (2007, 2012).

[23]We use worker's total pre-tax wage and salary income to calculate average hourly real wages (nominal values are adjusted by using the CPI-U adjustment factor to 1999 dollars (i.e., for the base survey year 2000)). There are two issues related to this: First, the CPS top-codes high wage incomes for reasons of confidentiality. This leads to an underestimation of wages in general and especially in the finance sector: Over all our survey years around 0.8% of workers in the X-sector are top-coded whereas in the Z_1-sector it affects around 1.6% of the workers and in the Z_2-sector even 7.6%. To dampen the bias in high wages we multiply top-coded incomes for survey years 1980-1995 by 1.5; a standard factor used in literature (as is described in Philippon and Reshef (2007, 2012)). From year 1996 on, top-coded wages are categorized into groups with different mean incomes by the CPS and thus the aggregate and the average wage income are uninfluenced by the top-coding. Note that the results are not very sensitive with respect to the multiplication factor (e.g., compared to $\omega = 1.62$, $\Psi = 5.08\%$ and $\Phi = 13.99\%$ in Table 2.2 resulting from factor 1.5, using a factor of 1.75 as in Philippon and Reshef (2007, 2012) would results in $\omega = 1.63$, $\Psi = 5.09\%$ and $\Phi = 14.02$). Second, a worker's total wage income consists of both wage income from longest job last year and wage income from other work. We cannot allocate these two incomes to different industries. Thus, we allocate the total wage income to one industry. If one assumes that the switch of job occurs equally likely between the three sectors, it does not bias the results. Furthermore, only about one fifth of all workers (in all three sectors) is affected by this; and of those who are affected not even a fourth of total income is coming from other work.

[24]Since the skill premium is approximately identical in all three sectors in the U.S. the skill intensities in the sectors need not be "normalized". They already correspond to the frictionless numbers.

[25]See Célérier and Vallée (2016) or Philippon and Reshef (2007, 2012) for a detailed empirical discussion of the finance premium.

spond to the "actual" numbers. The black lines correspond to the "normalized" ones.

2.8.1.2 Empirical trends

As is described in the introduction and picked up in the model, financialization has several aspects: On the one hand, the weight of the financial sector relative to non-financial business has increased; this is structural change towards finance. On the other hand, the type of financial products and services has changed; this is structural change within finance. The next two figures show the twofold structural change.

Figure 2.5: Employment ratio and wage sum ratio of the financial sector

Notes: Ψ^E measures the employment ratio (in terms of total hours worked) of finance (including insurance) compared to the rest of the U.S. economy. Ψ measures the ratio of the total wage sum in finance vs. the rest of the U.S. economy. "Actual" uses the observed sector-specific hours worked and hourly wages (for low-and high-skilled), whereas "normalized" uses the X-sector hours worked and hourly wages (for low-and high-skilled). Survey years from 1980-2013. Source: Own calculations based on CPS.

Figure 2.5 shows the ratio of the total finance sector (Z-sectors) compared to the non-finance economy (X-sector) for the U.S. based on the CPS data. On the one hand, the figure shows that finance has attracted new employment. The employment ratio (in terms of total hours worked) of the financial sector, defined by $\Psi^E_{actual} \equiv \frac{h^H_{z_1}\bar{H}_{z_1}+h^H_{z_2}\bar{H}_{z_2}+h^L_{z_1}\bar{L}_{z_1}+h^L_{z_2}\bar{L}_{z_2}}{h^H_x\bar{H}_x+h^L_x\bar{L}_x}$, increased from 4.54% in 1980 to 5.63% in 2013. The respective "normalized" ratio $\Psi^E_{normalized} \equiv \frac{h^H_x\bar{H}_{z_1}+h^H_x\bar{H}_{z_2}+h^L_x\bar{L}_{z_1}+h^L_x\bar{L}_{z_2}}{h^H_x\bar{H}_x+h^L_x\bar{L}_x}$ rose from 4.21% in 1980 to 5.18% in 2013. On the other hand, the figure illustrates the structural change towards the financial sector in terms of a growing wage sum ratio of finance. The wage sum ratio of the financial sector, defined as $\Psi_{actual} \equiv \frac{w^H_{z_1}h^H_{z_1}\bar{H}_{z_1}+w^H_{z_2}h^H_{z_2}\bar{H}_{z_2}+w^L_{z_1}h^L_{z_1}\bar{L}_{z_1}+w^L_{z_2}h^L_{z_2}\bar{L}_{z_2}}{w^H_x h^H_x\bar{H}_x+w^L_x h^L_x\bar{L}_x}$, increased by

50% from about 5.17% in 1980 to 7.83% in 2013. The respective "normalized" ratio $\Psi_{normalized} \equiv \frac{w_x^H h_x^H \bar{H}_{z_1} + w_x^H h_x^H \bar{H}_{z_2} + w_x^L h_x^L \bar{L}_{z_1} + w_x^L h_x^L \bar{L}_{z_2}}{w_x^H h_x^H \bar{H}_x + w_x^L h_x^L \bar{L}_x}$ rose by 34% from 4.36% in 1980 to 5.94% in 2013. The difference between the employment (E) ratio and the wage sum ratio is the result of different skill-intensities in the different sectors. By comparing the "normalized" black with the "actual" gray lines one sees a large difference between the two ratios of the wage sum: More than half of the increase in the ratio of the wage sum is the result of the frictions (i) and (ii). Yet, as the black line shows, there is still structural change towards finance if one controls for the two frictions. Comparison of the two black lines shows that the difference between the employment ratio and the wage sum ratio increased over time.

Figure 2.6: Employment ratio and wage sum ratio within the financial sector

Notes: Φ^E measures the employment ratio (in terms of total hours worked) of "new finance" compared to "traditional finance". Φ measures the ratio of the total wage sum in "new finance" vs. "traditional finance". "Actual" uses the sector-specific hours worked and hourly wages (for low-and high-skilled), whereas "normalized" uses the X-sector hours worked and hourly wages (for low-and high-skilled). Survey years from 1980-2013. Source: Own calculations based on CPS.

We observe a similar pattern for the within finance sectoral structure by splitting total finance up into subsectors Z_1 and Z_2. Figure 2.6 shows the employment ratio and the wage sum ratio of finance subsector Z_2 compared to the subsector Z_1 for the U.S. since the 1980s based on the CPS data set. "New finance" (subsector Z_2) grew strongly independent of the measure we use: The within employment ratio (in terms of total hours worked) of finance subsector Z_2, $\Phi^E_{actual} \equiv \frac{h^H_{z_2} \bar{H}_{z_2} + h^L_{z_2} \bar{L}_{z_2}}{h^H_{z_1} \bar{H}_{z_1} + h^L_{z_1} \bar{L}_{z_1}}$, more than doubled from about 8.63% in 1980 to 20.79% in 2013. The respective "normalized" ratio $\Phi^E_{normalized} \equiv \frac{h^H_x \bar{H}_{z_2} + h^L_x \bar{L}_{z_2}}{h^H_x \bar{H}_{z_1} + h^L_x \bar{L}_{z_1}}$ is very similar with a rise from 8.26% in 1980 to 19.77% in 2013. The within finance wage

sum ratio, defined by $\Phi_{actual} \equiv \frac{w^H_{z_2} h^H_{z_2} \bar{H}_{z_2} + w^L_{z_2} h^L_{z_2} \bar{L}_{z_2}}{w^H_{z_1} h^H_{z_1} \bar{H}_{z_1} + w^L_{z_1} h^L_{z_1} \bar{L}_{z_1}}$, increased dramatically from 11.75% in 1980 to 29.02% in 2013 peaking in survey 2009 at 40.18%. The respective "normalized" ratio $\Phi_{normalized} \equiv \frac{w^H_x h^H_x \bar{H}_{z_2} + w^L_x h^L_x \bar{L}_{z_2}}{w^H_x h^H_x \bar{H}_{z_1} + w^L_x h^L_x \bar{L}_{z_1}}$ rose from 9.17% in 1980 to 22.83% in 2013 with a peak in survey year 2009 of 29.91%. Hence, about two-thirds of the actual rise in the wage ratio of "new finance" cannot be assigned to frictions: They are also observed in the "normalized" data. The rest of the rise comes from friction (ii) (finance premium), which is particularly strong in the finance subsector Z_2.

As argued in the introduction financialization (with the twofold structural change) and inequality are two closely related topics. Figure 2.7 shows the development of the "normalized" skill premium calculated by $\omega = \frac{w^H_x}{w^L_x}$ for the U.S. since 1980, based on the CPS data. It increased from 1.55 in 1980 to 1.91 in 2013.[26] The time trend in ω illustrates that wage inequality increased over time. Nowadays high-skilled workers earn nearly double as much as low-skilled workers per hour. If one accounts in addition for the fact that high-skilled workers work more hours, the income inequality is even larger (e.g., 2.19 in 2013).

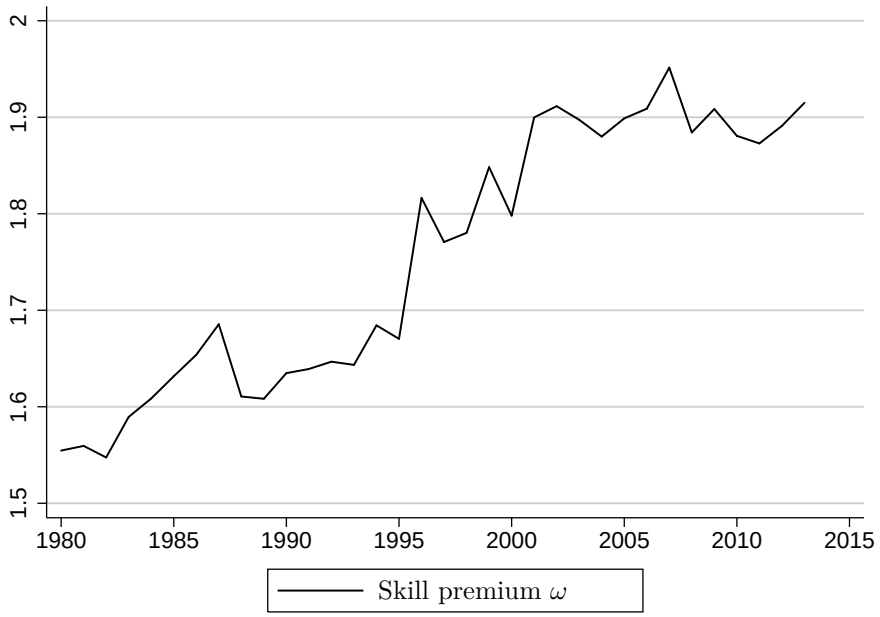

Figure 2.7: Skill premium

Notes: ω measures the "normalized" skill premium (i.e., hourly wage of high-skilled labor in X-sector divided by hourly-wage of low-skilled labor in X-sector). Survey years from 1980-2013. Source: Own calculations based on CPS.

[26]Interestingly, the skill premium in the U.S. is about the same in the three sectors because both low- and high-skilled workers in the financial industry earn a similar relative finance premium.

2.8.2 Numerics

In this section we implement our theoretical model quantitatively and use it for several numerical exercises. These illustrate possible drivers of the empirical developments presented in Figures 2.5-2.7. For the quantitative implementation of our model and the comparative-static equilibrium results we exclude the post-crisis years and compare average values for the period 1980-1994 with the respective average values for 1995-2009. More specifically: First, we calibrate our model for the average value of the early survey years 1980-1994. This calibrated model is then used for comparative static analysis. We introduce (i) ceteris paribus shocks and (ii) simultaneous shocks to illustrate how the channels analyzed in our model can generate the situation observed in later years (average values of later survey years 1995-2009).

2.8.2.1 Calibration

Table 2.1: Parameters survey years 1980-1994

Parameter	Data	Source	Description
$\bar{L}$	99.2m	CPS	# Low-skilled employees
$\bar{H}$	26.5m	CPS	# High-skilled employees
h^L	1639.4	CPS	Yearly hours of low-skilled
h^H	1982.6	CPS	Yearly hours of high-skilled
α_x	0.34	CPS	Output ela. of high-skilled in X
α_{z_1}	0.42	CPS	Output ela. of high-skilled in Z_1
α_{z_2}	0.68	CPS	Output ela. of high-skilled in Z_2
A_x	26.53	CPS	Technology level in X
PT_{65}	\$ 11,204	U.S. Bureau of the Census	Real poverty threshold <65
PT^{65}	\$ 10,076	U.S. Bureau of the Census	Real poverty threshold >65
$LEratio$	4.66	LE from World Bank	Old-age ratio
r^f	0.0368	Federal Reserve Bank of St.Louis	Real effective federal funds rate
A_{z_1}	116	Model calibration	Technology level in Z_1
A_{z_2}	165	Model calibration	Technology level in Z_2
δ	0.385	Model calibration	Discount rate
μ	0.740	Model calibration	Certainty measure

Notes: The table shows the averaged values for the time range of survey years $t \in \{1980, \ldots, 1994\}$. Averages of $\alpha_{j,t} = \frac{\kappa_{j,t}\omega_{j,t}}{1+\kappa_{j,t}\omega_{j,t}}$ with $\kappa_{j,t} = \frac{h^H_{j,t}\bar{H}_{j,t}}{h^L_{j,t}\bar{L}_{j,t}}$ and $\omega_{j,t} = \frac{w^H_{j,t}}{w^L_{j,t}}$, $j \in \{x, z_1, z_2\}$, $h^H_t = h^H_{x,t}$ and $h^L_t = h^L_{x,t}$. $A_{x,t} = \frac{w^L_{x,t}}{\Gamma_{x,t}\omega_{x,t}^{-\alpha_{x,t}}}$ with $\Gamma_{x,t} = {\alpha_{x,t}}^{\alpha_{x,t}}(1-\alpha_{x,t})^{1-\alpha_{x,t}}$. PT is the average, real poverty threshold of a two-people household (nominal values are adjusted by using the CPI-U adjustment factor to 1999 dollars (i.e., for the base survey year 2000) from CPS with PT_{65} denoting the relevant value for households younger than 65 and PT^{65} denoting the value relevant for older ones. $LEratio$ is the average ratio of working-time to retirement: $(65-20)/(LE_t - 65)$, where LE_t denotes life expectancy in year t; 65 is the retirement age and 20 is the assumed start of the working-life. r^f is the average, real effective federal funds rate (effective federal funds rate adjusted with the CPI-U adjustment factor from CPS). See bibliography for details on data sources.

We calibrate our model such that it fits the data for the average of the survey years 1980-1994 (i.e., years 1979-1993). Exogenous values from data are used for labor endowments $\bar{L}, \bar{H}, h^L, h^H$, output elasticities α_j, technology in the X-sector A_x, interest rate r^f and poverty thresholds (PT_{65} for young and PT^{65} for old households) as summarized in Table 2.1. For the subsistence levels we assume that each worker must cover over the life cycle half of a two-people household's poverty threshold. Further, we account for the fact that during the 1980-1994 time period the ratio of working-time to retirement was $LEratio = 4.66$ (i.e., we divide the poverty threshold of old households by 4.66). Hence, $\bar{e}_0 = PT_{65}/2$ and $\bar{e}_1 = PT^{65}/2/4.66$. The real safe return is $r = 1 + r^f$ with r^f being the real effective federal funds rate and the risky return is such that the risk premium is four percentage points (i.e., $R = (r + 0.04)/\mu$). We measure the efficiency units from the model by $b_l = h^l$, $l \in \{H, L\}$, where h^l are hours worked.

The other parameters (productivities in the finance sectors A_{z_1} and A_{z_2}, discount factor δ and completeness measure μ) are calibrated internally by targeting wage inequality ω, "normalized" ratios for the sectoral structure Ψ and Φ of the U.S. economy and the gross saving rate in the U.S. for the average of the survey years 1980-1994. The targeted values are shown in Table 2.2. More specifically, we solve the model numerically for possible parameter combinations of A_{z_1}, A_{z_2}, δ and μ and grid-search for the combination (see Table 2.1 for calibrated values) which minimizes the sum of the squared relative distances of the four model values from the corresponding data targets.[27] The comparison of the four model values generated by our calibrated model with the data outcomes is given in Table 2.2:

Table 2.2: Targets

Variables	Model	Data	Source	Description
ω^*	1.63	1.62	CPS	Skill premium
Ψ	5.08%	5.08%	CPS	Between sectoral structure
Φ	13.99%	13.99%	CPS	Within sectoral structure
saving rate	20.32%	20.30%	World Bank	Aggregate savings

Notes: ω^* is the equilibrium skill premium (per hour worked). Ψ corresponds to $\frac{p_{z_1}D + p_{z_2}F}{X}$ in the model and to $\Psi_{normalized}$ in the data. Φ corresponds to $\frac{p_{z_1}D}{p_{z_2}F}$ in the model and to $\Phi_{normalized}$ in the data. The saving rate is $(D + F)/W$ in the model and the share of aggregate savings in gross national income in the data, where aggregate savings (gross savings) is gross national income less total consumption, plus net transfers. See bibliography for details on data sources.

The calibrated model fits the targets fairly well. Further, the other equilibrium values

[27] For solving the model numerically, we use the demand functions in the goods and financial services markets to obtain the equilibrium values of X-, Z_1- and Z_2 as functions of ω (and exogenous parameters). Substituting these functions for X-, Z_1- and Z_2 in one of the labor market clearing conditions, we can solve for the equilibrium skill premium ω^*. (Then, at ω^*, the other labor market is also cleared.) From ω^* follow factor prices and prices of financial services, output levels and employment in the three sectors and the sectoral structure of the economy in a straightforward way.

following from the model are also very similar to the values observed in the CPS data (given in brackets). Hourly wages in our model are $w^H = \$\,19.3$ ($\$\,19.3$), $w^L = \$\,11.8$ ($\$\,11.9$) and the resulting prices are $p_{z_1} = 0.25$, $p_{z_2} = 0.19$.[28] Labor employments in total hours are $H_x = 49215$m (49215m), $L_x = 156043$m (156287m), $H_{z_1} = 2710$m (2794m), $L_{z_1} = 6113$m (6022m), $H_{z_2} = 614$m (635m), $L_{z_2} = 472$m (468m). For the skill intensities we get $\kappa_x = 0.32 < \kappa_{z_1} = 0.44 < \kappa_{z_2} = 1.30$ ($\kappa_x = 0.31 < \kappa_{z_1} = 0.43 < \kappa_{z_2} = 1.30$), which shows that the two finance subsectors are more skill intensive than the rest of the economy. These numbers suggest that the calibrated model matches the U.S. economy in the survey period 1980-1994 fairly well.

2.8.2.2 Numerical exercises

We show now how our calibrated model can predict the twofold structural change and the rising wage inequality between survey period 1980-1994 and survey period 1995-2009 as seen in Figures 2.5-2.7. To do so, we look at the predictions of our calibrated model if shocked by exogenous changes. Thereby, we apply the changes in the exogenous parameters of our model as observed in data. In other words, we use as shocks the average values of $\bar{L}$, $\bar{H}$, h^L, h^H, α_x, α_{z_1}, α_{z_1}, A_x, PT_{65}, PT^{65}, $LEratio$ and r^f for the time span of the survey years 1995-2009 instead of the ones for the time span of the survey years 1980-1994.[29] In addition, we also consider shocks on the internally calibrated parameter A_{z_1}, A_{z_2}, δ and μ.

Table 2.3: Comparative statics

	ω	Ψ	Φ
Uniform productivity progress A_j (income effect)	1.63	5.18%	14.92%
X-biased technical change A_x	1.64	6.15%	16.24%
Z_1-biased technical change A_{z_1}	1.63	4.36%	10.02%
Z_2-biased technical change A_{z_2}	1.63	4.91%	15.26%
Skill-biased technical change α_x, α_{z_1}, α_{z_2}	2.49	5.08%	13.68%
Higher subsistence requirement young $\bar{e_0}$	1.63	5.08%	13.99%
Higher subsistence requirement old $\bar{e_1}$	1.63	5.28%	13.17%
Increased skill supply k	1.21	4.96%	14.99%
Lower safe return r ($\frac{\bar{e_1}}{r}$-channel)	1.63	5.10%	13.91%
Lower relative return ρ	1.64	5.06%	15.32%
More completeness μ	1.64	4.99%	24.75%
Fall in δ	1.63	4.68%	13.61%

Notes: Ceteris paribus comparative-static effects.

[28] The magnitude of the financial services prices could be interpreted in the following way: A household has to pay the unit costs of financial intermediation, estimated by Philippon (2015) to be 0.015-0.02, during all his/hers "capital-accumulation" years (i.e., 15-times from 1980-1994 to 1995-2009).

[29] See Table 2.1 in Appendix B.4 for data of the average values for survey years 1995-2009 of $\bar{L}$, $\bar{H}$, h^L, h^H, α_x, α_{z_1}, α_{z_1}, A_x, PT_{65}, PT^{65}, $LEratio$ and r^f. For R we use again a constant risk premium of four percentage points.

As a first exercise, we introduce ceteris paribus shocks. This means that we apply each of the changes listed in Table 2.3 separately. For the exogenous parameters we apply observed changes; for the internally calibrated parameters potential changes. The quantitative effects of such ceteris paribus changes on the skill premium ω, on the between sectoral structure Ψ and the within structure Φ are summarized in Table 2.3.

By comparing Table 2.3 with Table 2.2 we see the magnitude of different effects. Uniform productivity progress A_j means that the productivities in all three sectors $j \in \{X, Z_1, Z_2\}$ grow at the same rate (i.e., $A^1_{z_i} = g_x A^0_{z_i}$, where $g_x = A^1_x/A^0_x$ is given by the observed average values of A^0_x from survey years 1880-1994 and of A^1_x from survey years 1995-2009). Consistent with Proposition 2.4-2.6 such a uniform productivity progress leads to the twofold structural change. Also (however, only visible at later digits) the skill premium increases. This is due to the income effect arising through the subsistence requirements $\bar{e_0} > 0$ and $\bar{e_1} > 0$. Sector-biased technical change means that only the respective sector's productivity grows, while the other two productivity levels are kept constant (as growth rate we use always the observable rate g_x). The comparative static effects of such a ceteris paribus shock are a combination of income and substitution effects. (Sector-specific) skill-biased technical change α_j, as observed in the data for $j \in \{X, Z_1, Z_2\}$, induces clearly an increase of the skill premium. An increase in skill supply $k = \frac{\bar{H}h^H}{\bar{L}h^L}$ reduces the skill premium and leads to within structural change because there are more high-skilled people who demand more finance subsector Z_2 services. Furthermore, a lower relative return ρ (induced by an increase of the risk premium by one percentage point) or more market completeness μ (by ten percentage points) raise the skill premium and make new financial services relatively more attractive compared to services for deposits. Finally, a fall in δ to 0.335, which leads to a lower saving rate close to 18.83% as observed on average for the time span of survey years 1995-2009, leads to smaller financial sectors.

As a second exercise, we shock our calibrated model with simultaneous shocks. This means, we shock our economy by using all the shocks in the exogenous parameters together (i.e., new average values of $\bar{H}$, $\bar{L}$, h^H, h^L, α_x, α_{z_1}, α_{z_1}, A_x, PT_{65}, PT^{65}, $LEratio$ and r^f for time span of survey years 1995-2009). Further, we assume uniform technological progress. This means, the productivities in the Z-sectors develop identical to the productivity in the X-sector. Discount parameter δ and completeness measure μ are held fixed at the calibrated values. With this procedure, we get a quantitative model prediction which can then be compared with the empirical development (see Table 2.4). Under simultaneous shocks our model predicts a rise in the skill premium ω from 1.63 to 1.86 and twofold structural change towards and within finance with a rise of Ψ from 5.08% to 5.21% and a rise of Φ from 13.99% to 15.02%.

Table 2.4: Predictions

Variables	Model	Data	Source	Description
ω^*	1.86	1.85	CPS	Skill premium
Ψ	5.21%	5.54%	CPS	Between sectoral structure
Φ	15.02%	23.41%	CPS	Within sectoral structure

Notes: ω^* is the equilibrium skill premium (per hour worked). Ψ corresponds to $\frac{p_{z_1}D+p_{z_2}F}{X}$ in the model and to $\Psi_{normalized}$ in the data. Φ corresponds to $\frac{p_{z_1}D}{p_{z_2}F}$ in the model and to $\Phi_{normalized}$ in the data.

Comparing the model values with data, we see that the simulated equilibrium values underestimate the between structural change (only a little) and mainly the within structural change. This means, additional shocks are needed to come closer to data values. According to our analysis, possible candidates for such additional shocks (unobserved in our data) are, for example, more market completeness (μ-shock shown in Table 2.3) or diminished fixed costs in the financial sector and distorted portfolio choices as discussed in Appendix B.2. Overall, the simulated development in our calibrated model illustrates the channels that lead to the observed rise in the skill premium and the twofold structural change towards and within the financial sector fairly well; at least as far as these changes are caused by economic fundamentals. As pointed out in the beginning of this section, the normalized financial sector ratios considered here are amplified in reality by rents.

2.9 Conclusion

The presented 3x3 model of production and financial services helps to explain the twofold structural change towards and within the financial sector. The analysis emphasized demand side effects by using quasi-homothetic preferences of the Stone-Geary form and accounted for supply side effects by considering for different skill-intensities in production of goods and financial services. The theoretical analysis was based on established building blocks for modeling a multi-sector economy with production and was at the same time sufficiently tractable to allow analytical results. The comparative-static equilibrium analysis showed the effects of productivity progress and technical change, skill supply, present and future subsistence requirements and financial product innovation on the skill premium and on the sectoral structure of an economy. Both the size of the financial sector relative to the non-financial sector as well as the size of the new finance sector relative to the traditional finance sector were considered. Moreover, in a supplementary appendix several extensions the robustness of the results was discussed and the effects of rents or distortions in the financial sector were addressed. The main insight of the results from the theoretical analysis can be summarized as follows: If one looks for a single economic source (apart from assuming rents or distortions) that could explain the twofold struc-

tural change towards and within finance and the rising skill premium simultaneously, the income effect is a robust candidate. Other channels, like relative price effects within the financial sector lead to more ambiguous results.

The qualitative results derived in the theoretical analysis were illustrated quantitatively by calibrating the model to U.S. data from 1980-1994. Focusing on normalized data, which exclude rents, the numerical implementation of the model shows that the subsequent development observed in the period 1995-2009 can be explained fairly well. While uniform productivity growth, working through the income effect, is confirmed as a main source of structural change towards and within finance, skill biased technical change is important too for matching the rise in the skill premium.

The paper leaves open two main questions which are important in the current debate about real economic development and financialization. The first open problem is the finance premium. While it is obvious that the rents revealed by the premium contribute to inequality and blow up the structural change towards and withing finance considered in this paper, the question where the premium comes from is less clear. In recent years, several attempts have been made to explain the premium by asymmetric information between shareholders and employees in the banking sector. Yet, this can only explain the redistribution of earnings within the financial sector. Our hypothesis is that it is the asymmetry between financial agents and their clients which allows to extract rents. After all, the financial sector is an expert system to start with. Possible channels for modeling the rent-generating information asymmetry would be intransparent cost structures or confusion by financial innovation (distorted μ-beliefs).

The second open question left to future research is how structural change towards and within the financial sector affects economic productivity. The literature on financial development and growth has identified market completion by financial innovation as an important source of growth. Does the recent evidence on a negative effect of financial development on economic growth indicate that the huge flood of new financial products since the 1990s has not really completed markets but rather generated obfuscation? In the framework presented in this paper such obfuscation would induce euphoric beliefs about the degree of market completeness (μ), which is one of the drivers of structural change within finance and at the same time a possible lever for rent extraction. Another possible channel for a growth dampening effect could be the absorption of high-skilled labor in the finance sector, which leads to scarcity of talent outside the financial sector and may slowdown productivity growth.

To take stocks: The empirical evidence shows that the expansion of the financial sector and the changing structure within the financial sector towards new finance are partly caused by the finance premium. This is a rent which remains unexplained in the presented

paper. But there are also economic fundamentals which drive the twofold structural change. These drivers are the focus of the paper. The main explanation for the observed twofold structural change is a rise in average income generated by uniform productivity growth across sectors and factors, which changes demand for financial services, combined with skill-biased technical change that drives up the skill premium.

Could the structural change towards and within finance, accompanied by a rise in the skill premium, come to a halt? According to our model, apart from a slowdown of growth, the following factors exert downward pressure on finance shares and skill premium: Finance-biased productivity progress, less attractive risky investments, a decline in the saving rate or a stop in the proliferation of new financial products.

3 Environmental policy and political stability in China

3.1 Introduction

China has experienced substantial economic growth and serious environmental deterioration in the past decades. China's environmental pollution surpasses other countries in their similar phase of economic development and industrial structure. More specifically, the current environmental situation can be summarized as follows: (i) Environmental pollution is severe. Newspapers and academic papers have documented, for example, the immense CO_2 emission, the smog and low air quality in major cities of China, etc. (ii) Local governments have low incentives to improve environment, due to career concerns and availability of fiscal budget (Wu et al., 2013). (iii) There are increasing cases of household protest due to environmental issues (Wang, 2010).

This research project analyzes China's environmental situation from a political economy perspective and raises two questions: First, what are potential determinants of environment policy in the Chinese political system? Second, how do these factors influence China's environmental quality and household welfare? For answering these questions, this paper presents a model with a central government, and local officials and households from N regions, which captures three important aspects of Chinese reality: The cadre system of the central government, households' protest against local officials, and the allocation of local fiscal revenues. With this framework we can evaluate the factors that influence environment and agents' behavior at different levels.

The main contribution of this paper is to provide a theoretical framework that allows to analyze the behavior of the central government, local officials and households. We evaluate the impact of their interaction on local environment and household welfare. The model consists of the following key elements: First, a benevolent central government chooses an environmental policy - a pollution abatement technology, and employs local officials to implement the technology. The central government is unable to observe local officials' ability directly, and infers their ability by observing local outcome. Specifically, we explicitly characterize the cadre system of the central government as a scheme of rewards contingent on economic output and household protest. Second, local officials

differ in their ability. They influence local production and implement environmental policy by allocating local fiscal revenues to production- and environment-related infrastructure, respectively. Third, households make protest decision according to their realized utility (from consumption and environment). Household protest and the central government's cadre system indirectly determine the investment allocation of local officials in production and in environment.

Within the theoretical framework we focus on the principle mechanisms. First, we model households' protest decision in the most simple way. Namely, households protest when their utility falls below an exogenous threshold. Second, local officials allocate fiscal revenues so that the induced outcome maximizes their own expected utility. Thus, the central government's reward scheme has a direct impact on the tradeoff of local officials in determining investment allocation. Furthermore, local officials' behavior is influenced by their ability and the central government's environmental policy. Lastly, the central government aims to maximize aggregate household welfare when choosing the environmental policy. However, it has to account for the local officials' behavior and possible household protests. Specifically, we assume in the model that local household protest *per se* incurs no costs. Nevertheless, when the share of regions with households who protest exceeds a certain threshold, the protest generates negative externalities on the entire population. We regard this situation as political instability. Households do not internalize the externalities when deciding protest. However, the central government needs to consider the impact of political instability on household welfare when choosing the optimal environmental policy.

We characterize the central government's environmental policy and local officials' investment allocation in the equilibrium emerging under the cadre system of the central government and compare the outcome with a social optimum benchmark. In contrast to the equilibrium with opportunistic local officials, in the considered social optimum benchmark, local officials are benevolent and maximize the expected utility of households from their region.[1] Therefore, the local officials' allocation of fiscal revenues in the social optimum deviates from the one in equilibrium (with opportunistic local officials).

We quantitatively solve for the optimal environmental policy of the central government, and the optimal investment allocation of local officials under the two cases - social optimum and equilibrium, respectively. The findings are intuitive. Under social optimum, local officials balance household benefits from investment in production and in environment. The central government chooses the optimal environmental policy to avoid political instability. In equilibrium under the central cadre system, local officials overinvest in pro-

[1]The social optimum here is, however, not first best, because potential externalities from political instability are not taken into account at the local level.

duction to increase their probability of promotion. As a result, environment deteriorates. This is not only the case in regions with low development and low ability officials. On the contrary, the marginal impact of investment on promotion probability is higher for high ability officials. Therefore, they have more incentive to produce at the cost of environment, so that the environmental quality in high ability regions is low. Household protest constrains local officials' investment in production-related infrastructure. This is especially the case in medium-ability officials, where the probability of promotion is low, and thus the income loss due to household protest and demotion dominates. As a consequence, production is restricted and environment is good. Finally, under the cadre system the central government chooses a policy less effective than the socially optimal one, in order to avoid the negative externalities by household protest and overproduction by local officials in pursuit of promotion.

Lastly, the quantitative results are used to evaluate the welfare effect of the central government's policy and local officials' behavior, and their impact on environment. The results are in line with the empirical observation of China's "high output, high pollution" situation. First, due to the impact of the cadre system on local officials' behavior, households from all regions consume more than under thesocial optimum. But the environmental quality is lower. This happens especially in regions with high-ability officials. However, environment deterioration is not only the result of inadequate environment-related investment at the local level. Also the pollution abatement technology chosen by the central government is laxer than under the social optimum.

3.1.1 Related literature

This research project is related to several strands of research.

The main objective of the paper is to explain China's "high output, high pollution" mix from a political economy perspective.[2] This is related to papers that analyze how political institutions influence environmental policy and officials' behavior, and determines the environment of the economy.[3] To the best of my knowledge, none of the previous studies simultaneously considers the determination of environmental policy, local officials' behavior, and household protest, as well as how these behavior influences environmental quality and household welfare.

This paper explicitly characterizes the cadre system of the central government, and

[2]See Zheng and Kahn (2013) and Chang et al. (2015) for a description of the status quo of China's environment.

[3]Jia (2014) studies how connection with top officials in the central government influences local officials' behavior and the regional environmental quality. Fredriksson and Wollscheid (2014) discuss the impact of political institutions on stringency of environmental policy.

analyzes its impact on local officials' behavior.[4] Thereby, the central government and local officials are in a hierarchical relationship (Maskin et al., 2000), and is closer to a principle-agent relation as in industrial organization: The central government sets evaluation standards, and local officials are motivated by career concerns. We model local officials' production and promotion incentives as in Holmström (1999) and Alesina and Tabellini (2007).

Furthermore, this paper embeds local officials' allocation of fiscal revenues. Local government in China in general has the authority to decide about the allocation of fiscal revenues.[5] Their decision on the allocation has crucial impact on the implementation of the central government's policy and on local development. Wu et al. (2013) show empirically that local officials overinvest in GDP-related infrastructure. However, theoretical modeling is so far still missing.

Finally, this paper is related to the broad research on government accountability. List and Sturm (2006), Besley and Case (1995) and Besley and Burgess (2002) analyze in a democratic economy determinants of government accountability. They emphasize that reelection motives influence officials' behavior and their policy, and point out that voters supervise accountability of the government. In contrast, households are not directly involved in election of government officials in China. Moreover, the central government has relatively weak influence on local officials' decision about environment-related investment (Saich, 2012). This paper considers the accountability of local officials in the context of China from a new perspective: The impact of household protest on the decision of governments, especially on the investment allocation of local officials.[6] This channel gains importance along with the popularization of internet and the emergence of environmental NGOs; it becomes easier and less costly to coordinate among households in protest. Therefore, one could imagine an increasing impact of household protest on the accountability of local officials in China.

The rest of the paper is structured as follows: In section 3.2 the theoretical model is described. Section 3.3 and 3.4 characterize household protest decision, central environ-

[4]See Edin (2003), Li (1998), Li and Zhou (2005), Su et al. (2012), Rochlitz et al. (2014), and Xu (2011) for a detailed discussion of the Chinese cadre system.

[5]See Tsui and Wang (2004) and Jin et al. (2005) for a description of the decentralized fiscal system in China. See Asian Development Bank (2014) for a discussion of China's local fiscal management.

[6]The impact of household protest on government in China is documented in many papers. Saich (2012) gives a comprehensive discussion of households' protest and the influence on policy implementation. Using survey data, the paper additionally demonstrates that household protest in China happens mainly at local level and households in general believe that the central government is benevolent. This is reflected in the data: In 2011, the share of households (respondents) who are satisfied with the central government is 91.8%, whereas this share drops to 63.8% at local level. From this aspect, it makes sense to assume a benevolent central government with career-concern local officials in the model to characterize the political agents in China.

mental policy and local investment allocation for the social optimum benchmark and the equilibrium under the cadre system, respectively. The quantitative results are illustrated and discussed in Section 3.5. Section 3.6 concludes.

3.2 The model

Consider a one-period economy with a central government and a mass N of isolated regions. The central government chooses an environmental policy by deciding about the pollution abatement technology to be implemented. At the same time, they take charge of the assignment and evaluation of local officials. In each region there is one local official and a mass 1 of households. Local officials supervise local production, collect tax revenues and implement the environmental policy. Specifically, they allocate the tax revenues to production- and to environment-related investments to enhance output and probability of successful policy implementation, respectively. Production generates pollution, whereas successful environmental policy implementation reduces the level of pollution locally. Households from all regions are identical. They derive utility from consumption and good environment. The regions only differentiate with respect to the ability of their local official. With slight abuse of notation we use local officials' ability, a, as an indicator of the different regions in the economy.

3.2.1 The central government

The central government is benevolent. It maximizes the aggregate expected utility of all households when choosing the environmental policy. In addition, it assigns and evaluates local officials.

3.2.1.1 The central government's environmental policy

Suppose a linear environmental quality determination: $E = \bar{E} - \varphi Y$, where $\bar{E}$ is environmental quality in all regions at the beginning of the period, Y is local output, and φ is pollution per unit of output. We characterize an environmental policy as the decision to implement in each region a pollution abatement technology to reduce pollution in the production process. Suppose that there is a continuum set of technologies, $\theta \in [0, 1]$, that differ in their strength. If there is no technology ($\theta = 0$), or a technology is unsuccessfully implemented, pollution per unit of output is $\varphi = \bar{\varphi}$. Successful technology implementation reduces pollution per unit of output by $\varphi(\theta)$, and thus $\varphi = \bar{\varphi} - \varphi(\theta)$. Assume that

$$\varphi(0) = 0, \ \frac{\partial \varphi(\theta)}{\partial \theta} > 0, \text{ and } \frac{\partial^2 \varphi(\theta)}{\partial \theta^2} < 0. \tag{3.1}$$

Larger θ implies more effective abatement technology, whereas the marginal reduction in pollution decreases as the strength increases.

3.2.1.2 Cadre system for local officials

The central government employs local officials from a natural pool. The probability distribution of local officials' ability in the pool is $\Psi(a)$, where ability, a, is a random variable, $a \in [a_{min}, a_{max}]$. Assume that drawing people from the pool does not influence the ability distribution, $\Psi(a)$. The central government intends to select local officials with high abilities. However, local officials' ability is private information; the central government is unable to observe their ability directly nor indirectly through local environmental outcomes. We assume that the central government may observe the output level of each region Y, subject to some noise, $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Denote the observed output by $\hat{Y}$, with $\hat{Y} = Y + \epsilon$. In addition, they recognize it when households protest against their local official. Accordingly, they evaluate local officials and reward them contingent on the evaluation. The reward scheme comprises three parts: Promotion, stay in office, and demotion.[7] If the observed local output is above an exogenous threshold, $\hat{Y} > Y^*$, the local official will be promoted. However, if households from a region protest against their local official, the official is demoted and the central government employs a new official randomly from the pool. If neither of the two cases occur, the local official stays in office. The local officials' income in the three cases are $\left\{\bar{A}, M, \underline{A}\right\}$, respectively. $\bar{A} > M > \underline{A}$.[8]

3.2.2 Local officials

3.2.2.1 Ability and policy implementation

Local output, $Y = Y(a, I_1)$, is determined by the local official's ability, a, and local investment in production-related infrastructure, I_1. Assume that $\frac{\partial Y}{\partial a} > 0$, $\frac{\partial Y}{\partial I_1} > 0$, $\frac{\partial^2 Y}{\partial a^2} < 0$, $\frac{\partial^2 Y}{\partial I_1^2} < 0$. In addition, local officials can spend I_2 on environment-related investment. Given the central government's pollution abatement technology, θ, local officials successfully implement the technology and reduce pollution with probability $\pi(a, I_2; \theta)$, and fail with probability $1 - \pi(a, I_2; \theta)$. We label the two states as high and low, $\{H, L\}$, respectively. The probability distribution depends on local officials' ability (e.g., the ability to coordinate and organize the policy implementation), local investment in environmental-related infrastructure, I_2, and the complexity of the task, θ.

[7] See Edin (2003) and Xu (2011) for more detailed description of the Chinese cadre system.

[8] The reward scheme and the threshold of promotion are exogenous in the model. It is for future research to investigate the question of optimal reward and promotion scheme.

Assumption 3.1 (Probability of policy success)**.**

$$\frac{\partial \pi(a, I_2; \theta)}{\partial \theta} < 0, \ \frac{\partial \pi(a, I_2; \theta)}{\partial a} > 0, \ \frac{\partial \pi(a, I_2; \theta)}{\partial I_2} > 0, \ \textit{and} \ \frac{\partial^2 \pi(a, I_2; \theta)}{\partial a \partial I_2} > 0.$$

The first three conditions are straightforward: More effective policies are less likely to succeed for all officials. Both ability and investment in infrastructure increases the probability of policy success. The last condition implies that ability and investment are complementary: A higher investment increases the marginal impact of ability on the probability of policy success, and vice versa.

3.2.2.2 Local environmental quality

The environmental quality in the two states are given respectively by

$$E^s = \begin{cases} \bar{E} - (\bar{\varphi} - \varphi(\theta))Y, & \text{if } s = H, \\ \bar{E} - \bar{\varphi} Y, & \text{if } s = L. \end{cases} \tag{3.2}$$

Notice that in a good state the environmental quality depends on local output level, Y, and the environmental policy, θ. Whereas in a bad state, it only depends on local output. In other words, local officials' investment in environment-related infrastructure influences the local expected environmental quality, but not the quality in high and in low state *per se*.[9]

Using (3.2), we can define the level of production sustainable by the environment in high and in low states, $Y^s_{sus}, s \in \{H, L\}$.

Definition 3.1 (Sustainable production)**.** *A production level Y in state $s \in \{H, L\}$ is sustainable by the environment if the environmental quality is non-negative. Namely, $E^s \geq 0$, where E^s is defined in* (3.2)*.*

Therefore, the maximal sustainable production in high and in low states are given respectively by

$$Y^H_{sus} = \bar{E}/(\bar{\varphi} - \varphi(\theta)) \text{ and } Y^L_{sus} = \bar{E}/\bar{\varphi}. \tag{3.3}$$

In particular, one should notice that $Y^H_{sus} > Y^L_{sus}$. In other words, output level that is sustainable in a high state may not be sustainable in a low state. This is crucial in

[9]This set-up is different from related paper (e.g. Jia (2014)) that assume officials choosing between clean and dirty technologies to produce, which directly influence the environmental outcome. However, since local officials' investment in environment-related infrastructure influences the expected environmental quality, a reallocation of investments from production to environment leads to similar outcome as to produce with clean technology. Assuming local officials influencing the probability of the states gives us technical simplicity in calculating the share of household protest and the optimal environmental policy.

formulating the central government's environmental policy decision in the equilibrium under the cadre system in Section 3.4.2.

3.2.2.3 Local investment and production

Local infrastructure investments, I_1 and I_2, are financed by local tax revenues. The tax revenues, Π, come from a proportional tax on output,

$$\Pi = \tau Y(a, I_1), \tag{3.4}$$

where the tax rate τ is exogenous. Given local officials' budget (3.4), we can define the maximal feasible investment.

Definition 3.2 (Maximal feasible investment)**.** *The maximal feasible investment is the maximal amount of investment a local official with ability a can make.*

Formally, the maximal feasible investment is determined as follows:

$$\begin{aligned} \max_{I_1, I_2} \quad & \tau Y(a, I_1) \\ \text{s.t.} \quad & I_1 + I_2 \leq \tau Y(a, I_1) \end{aligned}$$

This gives us directly the following property.

Property 1. *The maximal feasible investment equals to the maximal tax revenues a local official may collect. It is achieved when local officials invest all tax revenues into production-related infrastructure. Namely,*

$$I_1^{max} = \tau Y(a, I_1^{max}) \tag{3.5}$$

From now on, we use $I_1^{max}(a)$ to denote the maximal feasible investment of local officials with ability a. At $I_1^{max}(a)$, local output also achieves its maximal, $Y^{max}(a)$, and investment in environment-related infrastructure is zero. Apparently, both $I_1^{max}(a)$ and $Y^{max}(a)$ increase in local officials' ability a.

Local officials derive utility from the income rewarded by the central government contingent on evaluation. Given the reward and evaluation scheme of the central government, local officials decide the optimal allocation of investments, subject to local fiscal budget constraint, to maximize their expected utility.

3.2.3 Households

Households derive utility from consumption and good environment. The instantaneous utility function is given by $U(C, E)$. It satisfies Inada conditions on $E \geq 0$, with $U(0, E) = U(C, 0) = 0$.[10] And on $E < 0$, in which case production is beyond the sustainable threshold defined in Definition 3.1, household utility is negative infinity. We simplify the mircofoundation of how households acquire income (e.g., through labor and capital supply in local firms), and set their consumption to be the net output of the region, $(1 - \tau)Y(a, I_1)$.

Households may protest against their local official if their realized utility falls below a threshold, $\underline{U}$. In case of protests, the central government will replace the local official with a new one randomly drawn from the natural pool.[11]

Protest incurs zero cost *per se*. However, when the share of households who protest in the economy exceeds a threshold, $\bar{\lambda}$, protest behavior generates negative externalities on the entire population.[12] Households do not consider the externalities when making protest decision, but the central government takes this into account when deciding about the environmental policy. For simplicity, we assume household utility is zero when large scale (i.e., above $\bar{\lambda}$) protests occur.

3.2.3.1 Thresholds of household protest

Households' utility is determined by local output and environmental quality. Environmental quality in turn depends on local output. Furthermore, depending on whether the environmental policy is successfully implemented, the realization of local environmental quality can be high or low, defined in (3.2). Therefore, given local output and the state of the environment, household utility is uniquely determined.

To simplify the further discussion, we redefine household utility in state $s \in \{H, L\}$, as a function of local output Y. Specifically,

$$U^s(Y; \theta) \equiv U(C, E^s) = U\left((1 - \tau)Y, \bar{E} - \left(\bar{\varphi} - \mathbb{1}_{\{s=H\}}\varphi(\theta)\right)Y\right), \tag{3.6}$$

[10] Namely, $\frac{\partial U(C,E)}{\partial C} > 0$, $\frac{\partial U(C,E)}{\partial E} > 0$, $\frac{\partial^2 U(C,E)}{\partial C^2} < 0$, $\frac{\partial^2 U(C,E)}{\partial E^2} < 0$, $\lim_{C\to 0} \frac{\partial U(C,E)}{\partial C} = \infty$, $\lim_{C\to\infty} \frac{\partial U(C,E)}{\partial C} = 0$, $\lim_{E\to 0} \frac{\partial U(C,E)}{\partial E} = \infty$, $\lim_{E\to\infty} \frac{\partial U(C,E)}{\partial E} = 0$.

[11] We fix the threshold of household protest as constant and exogenous in the model. Yet, it can be easily extended to the case when protest threshold depends on the level of local production (e.g., $\underline{U} = U(Y)$, $\frac{\partial U}{\partial Y} > 0$). In addition, we could simplify the behavior of local officials and solve for the threshold of household protests endogenously. For example, households compare the expected utility if sticking to the incumbent, which is calculated with households' belief on the official's ability, versus protesting and getting a new official with average ability in the pool. In this case, an endogenous threshold of households' belief on local officials' ability, below which households will protest, emerges.

[12] In reality, one could think that a large share of regions with household protest induces political instability, both of which influence household welfare of the entire population negatively.

where $s \in \{H, L\}$, and $\mathbb{1}_{\{s=H\}}$ is an indicator function that equals one when the high state is realized, and pollution is reduced. Following the definition, the threshold of household protest, $\underline{U}$, can be mapped into thresholds of output levels. First, define the maximal achievable household utility in state $s \in \{H, L\}$, U^s_{max}, as

$$U^s_{max} = \max_{Y} \quad U^s(Y; \theta). \tag{3.7}$$

Apparently, $U^H_{max} > U^L_{max}$. Now, if $\underline{U} > U^H_{max}$, household utility is lower than the threshold regardless of the realized state and their local official's ability and investment allocation. Therefore, households always protest. If $U^L_{max} < \underline{U} \leq U^H_{max}$, households always protest in a bad state by the same argument. In a good state they protest only if their utility falls below the threshold. This is the case when local output is low or overly high. Intuitively, both low consumption (due to low output), and bad environmental quality (due to overly high production) result in low utility, which induces household protest. In the end, if $\underline{U} \leq U^L_{max}$, households may protest in both good and bad state for similar reason as above. We summarize the result in the following lemma.

Lemma 3.1 (Household protest)**.**

i) If $\underline{U} > U^H_{max}$, households will protest regardless of the behavior of their local official.

ii) If $U^L_{max} \leq \underline{U} \leq U^H_{max}$, households always protest if a low state is realized. If a high state is realized, there exists an interval $[\underline{Y}^H, \bar{Y}^H]$, such that households protest if and only if local output is outside the interval.

iii) If $\underline{U} < U^L_{max}$, for each state $s \in \{H, L\}$, there exist an interval, $[\underline{Y}^s, \bar{Y}^s]$ such that households protest if and only if local output is outside the interval $[\underline{Y}^s, \bar{Y}^s]$ when state s is realized.

In particular, $\left\{\underline{Y}^s, \bar{Y}^s\right\}_{s\in\{H,L\}}$ are solutions to the equation $U^s(Y; \theta) = \underline{U}$, where U^s is defined in (3.6)*. $\underline{Y}^H \leq \underline{Y}^L < \bar{Y}^L \leq \bar{Y}^H$, where equalities are achieved when $\theta = 0$.*

Proof. See text and illustration below. □

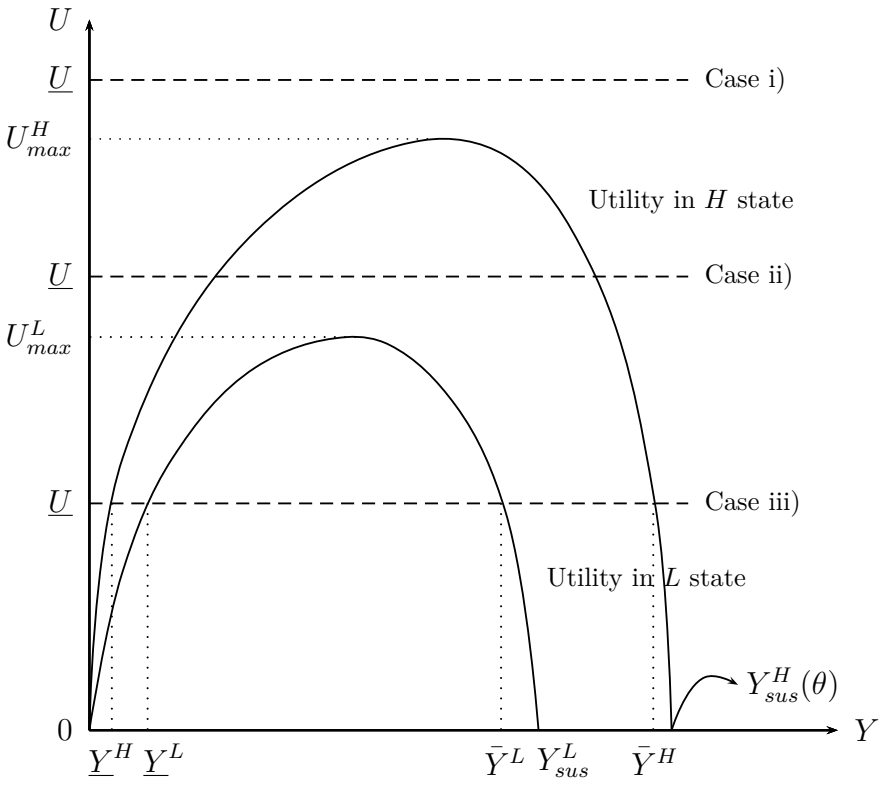

Figure 3.1: Illustration of output thresholds

We illustrate the three cases and the thresholds of case iii) in Figure 3.1. Since household utility satisfies Inada conditions, the curve displays an inverse U-shape as output increases. In addition, at a given level of output, household utility is higher when environmental quality is better. Therefore, household utility in high state, U^H, is above their utility in low state, U^L. Clearly, if local output is outside the interval $[\underline{Y}^H, \bar{Y}^H]$ in high state, or outside the interval $[\underline{Y}^L, \bar{Y}^L]$ in low state, household utility falls below the threshold, $\underline{U}$, and they protest. In addition, the upper interception of household utility in high and in low states with the horizontal Y-axis are the corresponding maximal sustainable production $Y^s_{sus}, s \in \{H, L\}$ defined in (3.3). Output beyond these levels results in negative infinite household utility.

In the following discussion, we focus on case iii), which induces a more comprehensive analysis of investment allocation in the equilibrium under the cadre system in Section 3.4. Technically, case i) and case ii) are special cases of case iii).

Finally, to simplify further analysis, we make the following assumption on the lower bound of local officials' ability, a_{min}.

Assumption 3.2 (Lower bound of local officials' ability)**.** *$a_{min} > \underline{a}^L$, where $\underline{a}^L$ is the minimal ability level such that output $\underline{Y}^L$ is achievable, $Y_{max}(\underline{a}^L) = \underline{Y}^L$.*

Clearly, the assumption excludes local officials against whom households always protest, regardless of their investment allocation. In addition, we will see in Section 3.3.1 and 3.4.1 that the assumption implies that the output in all regions - under social optimum and in the equilibrium under the cadre system - is always above $\underline{Y}^L$. Namely, in the end household only protest due to deteriorated environment (resulted from production above $\bar{Y}^s, s \in \{H, L\}$), rather than lack of consumption.

3.2.4 Model timing

We consider a one-period static model. At the beginning of the period, the central government chooses an environmental policy, θ. Then local officials decide the allocation of tax revenues on production-related and environment-related infrastructure investments, I_1 and I_2, and produce output, $Y(a, I_1)$. Production generates pollution, $\bar{\varphi}Y(a, I_1)$. Depending on local officials' ability, a, and their investment in environmental-related infrastructure, I_2, the abatement technology will reduce the marginal pollution level by $\varphi(\theta)$ with probability $\pi(a, I_2; \theta)$. In the end, households make their protest decision according to whether their realized utility falls below the threshold, $\underline{U}$. Notice that local production, $Y(a, I_1)$, takes place before the environment state is realized. This means, households' consumption, $C = (1-\tau)Y(a, I_1)$, are the same irrespective of the state realization. Utility in high and low states differ only due to different environmental quality defined in (3.2).

3.3 Socially optimal policy, investment allocation and household welfare

Before characterizing the optimal choices of the central government, local officials and households in the equilibrium under the cadre system, we first derive the optimal environmental policy, investment allocation and the corresponding household utility for the social optimum benchmark with benevolent (rather than opportunistic) local officials.[13]

Definition 3.3 (Social optimum)**.** *An environmental policy and investment allocation is socially optimal, if local officials maximize the expected utility of households in their region, and the central government maximizes the expected aggregate utility of all households in*

[13]Notice that the social optimum here is not first-best, because local officials maximize expected utility of households in their region, without taking into account the negative externalities of household protest on the welfare of households from other regions. In other words, we simply analyze the case when local officials are benevolent to their own households. It is technically complex in a N-region model to consider coordination of local officials to eliminate externalities of household protest.

the economy, given the ability distribution, $\Psi(a)$, investment decision of local officials and protest behavior of households.

3.3.1 Local officials' optimal investment allocation

Households' expected utility in region a is given by

$$\mathbb{E}U(I_1, I_2; a, \theta) = \pi(a, I_2; \theta)U^H(Y; \theta) + (1 - \pi(a, I_2; \theta))U^L(Y; \theta), \tag{3.8}$$

where $Y = Y(a, I_1)$ is local output, and $U^s(Y; \theta), s \in \{H, L\}$, is defined in (3.6). Therefore, in the social optimum, a local official with ability a solves the following allocation problem:

$$\begin{aligned} \max_{I_1, I_2} \quad & \mathbb{E}U(I_1, I_2; a, \theta) \\ \text{s.t.} \quad & I_1 + I_2 \leq \tau Y(a, I_1). \end{aligned} \tag{3.9}$$

Local officials maximize household utility by equalizing the relative marginal return (in terms of households expected utility) of investment in production- and in environment-related infrastructure, I_1 and I_2, to the relative marginal cost of the two investments. We derive the first order condition of the program and illustrate the determination of the optimal solution in Appendix C.1.1.

The optimization problem (3.9) gives the optimal investment allocation, $\{I_1^*(a, \theta), I_2^*(a, \theta)\}$, as a function of local officials' ability, a, and central government's environmental policy, θ. Correspondingly, household utility in high and low states, $U^H(a, \theta)$ and $U^L(a, \theta)$, respectively, and the expected utility, $\mathbb{E}U(a, \theta)$ are determined.[14] We calculate the optimal solution numerically in Section 3.5.

Apparently, it is never optimal for local officials to produce below $\underline{Y}^L$ under Assumption 3.2, since their ability always allows them to produce at higher levels that give households higher utility (and this is consistent with the objective of benevolent local officials).

3.3.2 Central government's optimal environmental policy

Given the optimal investment allocation of local officials, the central government calculates the share of household protest, λ, households' expected utility, $\mathbb{E}U(a, \theta)$, and chooses the optimal environmental policy so that aggregate expected utility is maximal. In both the

[14]Notice that in a bad state the environmental policy, θ, does not influence household utility directly, but indirectly through local officials' investment allocation, $I_1^*(a, \theta)$.

good and the bad state, households protest when their utility falls below the threshold, $\underline{U}$. Therefore, the share of households who protest is given by

$$\lambda(\theta) = \int_{a_{min}}^{a_{max}} \left[\pi(a,\theta)\mathbb{1}_{\{U^H(a,\theta)<\underline{U}\}} + (1-\pi(a,\theta))\mathbb{1}_{\{U^L(a,\theta)<\underline{U}\}}\right] d\Psi(a), \tag{3.10}$$

where $\pi(a,\theta) = \pi(a, I_2^*(a,\theta);\theta)$, and $\mathbb{1}_{\{U^s(a,\theta)<\underline{U}\}}, s \in \{H,L\}$, is an indicator function that equals to one if household utility in state s is below $\underline{U}$ and thus they protest.

The central government's maximization problem is

$$\max_{\theta} \quad \mathbb{1}_{\left\{\lambda(\theta)\leq\bar{\lambda}\right\}} \int_N \mathbb{E}U(a,\theta) d\Psi(a), \tag{3.11}$$

where $\mathbb{1}_{\left\{\lambda(\theta)\leq\bar{\lambda}\right\}}$ equals to one if the share of household protest is below the threshold, $\bar{\lambda}$. When the share is above the threshold, negative externalities from political instability imply zero utility for all households. Apparently, this is undesirable as long as the economy has the possibility of ending up in political stability (in which cases households have non-negative utility). Therefore, we can rewrite the central government's problem defined in (3.11) as an aggregate utility optimization problem with a political stability constraint:

$$\begin{aligned} &\max_{\theta} \quad \int_N \mathbb{E}U(a,\theta) d\Psi(a), \\ &\text{s.t.} \quad \lambda(\theta) \leq \bar{\lambda}. \end{aligned} \tag{3.12}$$

3.4 Equilibrium under the cadre system

Compared with the social optimum benchmark, the one thing that changes in equilibrium under the cadre system is the local officials' investment behavior. Local officials are now incentivized by the cadre system of the central government, and decide about the allocation of tax revenues on production- and environment-related investments in an opportunistic way. That is, they maximize their own expected income rather than the expected utility of the households living in their region. In this section, we discuss the determination of local officials' investment allocation and the conditions that defines the central government's environmental policy. The quantitative results are given in Section 3.5.

3.4.1 Local officials' equilibrium investment allocation

In contrast to the socially optimal case, local officials maximize their own expected utility in equilibrium. As a result, the central government's cadre system and the corresponding reward scheme, which determines local officials' expected utility under different circum-

stances, have a crucial impact on local officials' behavior. Specifically, the cadre system consists of three parts: Promotion, stay in office and demotion, with $\bar{A} > M > \underline{A}$ as respective remuneration levels.

3.4.1.1 Demotion

Local officials are demoted if and only if households from their region protest against them. And household protest occurs when local output is outside the interval $[\underline{Y}^s, \bar{Y}^s], s \in \{H, L\}$ (see Lemma 3.1 case iii)). Therefore, the probability of demotion, π^D, of local officials with ability, a, is given by

$$\pi^D \equiv \pi(a, I_2; \theta) prob(Y \notin [\underline{Y}^H, \bar{Y}^H]) + (1 - \pi(a, I_2; \theta)) prob(Y \notin [\underline{Y}^L, \bar{Y}^L]). \tag{3.13}$$

Notice that output is not a random variable; given local officials' investment allocation, output Y is determined. Thus, $prob(Y \notin [\underline{Y}^s, \bar{Y}^s]), s \in \{H, L\}$ is either zero or one. In addition, since $[\underline{Y}^L, \bar{Y}^L] \subseteq [\underline{Y}^H, \bar{Y}^H]$, if $prob(Y \notin [\underline{Y}^H, \bar{Y}^H]) = 1$, then $prob(Y \notin [\underline{Y}^L, \bar{Y}^L]) = 1$. Intuitively, households are more likely to protest in a low state (due to worse environmental quality) compared to a high state. Therefore, if households protest in a high state, given the output level, they will also protest if a low state is realized.

3.4.1.2 Promotion and stay in office

If protest does not occur, local officials may either be promoted or stay in office with the current income. Specifically, local officials are promoted if the observed production, $\hat{Y} = Y + \epsilon$, is above the promotion threshold, Y^*, and stay in office otherwise. Therefore, the conditional probability of promotion is given by

$$\pi^P = prob(\hat{Y} > Y^*) = prob(\epsilon > Y^* - Y) = 1 - \Phi(Y^* - Y), \tag{3.14}$$

where $\Phi(\cdot)$ is the cumulative distribution function of $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Correspondingly, the conditional probability of stay in office is $1 - \pi^P$.

Now we can define the expected utility of local officials (LO), which is the sum of the expected reward in the three cases:

$$U^{LO}(I_1, I_2, a, \theta) = \pi^D \underline{A} + (1 - \pi^D) \left[\pi^P \bar{A} + (1 - \pi^P) M\right], \tag{3.15}$$

where the probability of demotion, π^D, and the conditional probability of promotion, π^P, are defined in (3.13) and (3.14), respectively.

Local officials choose the allocation of tax revenues on I_1 and I_2 in such a way that their expected utility under the cadre system defined in (3.15) is maximized. Moreover,

their investments are physically constrained by the maximal feasible investment, $I_1^{max}(a)$, defined in (3.5). The following proposition characterizes local officials' optimal investment allocation:

Proposition 3.1 (Optimal investment allocation)**.** *There exists a threshold of ability, $\underline{a}$, defined by $Y(\underline{a}, I_1^{max}(\underline{a})) = \bar{Y}^L$, such that*

i) For officials with ability $a \in [a_{min}, \underline{a})$, investment in production-related infrastructure is equal to the maximal feasible investment, $I_1^{max}(a)$; investment in environment is zero.

ii) Officials with ability $a \in [\underline{a}, a_{max}]$ produce output, $Y \in [\bar{Y}^L, \bar{Y}^H]$, where $\bar{Y}^s, s \in \{H, L\}$, is given in Lemma 3.1. As a consequence, according to (3.13)*, the probability of demotion $\pi^D = 1 - \pi(a, I_2; \theta)$, and the optimal allocation is thus determined by*

$$\begin{aligned} \max_{\{I_1, I_2\}} \quad & U^{LO}(I_1, I_2, a, \theta), \\ s.t. \quad & I_1 + I_2 \leq \tau Y(a, I_1), \\ & U^{LO}(I_1, I_2, a, \theta) \geq \Gamma(\bar{Y}^L), \end{aligned} \tag{3.16}$$

where $U^{LO}(I_1, I_2, a, \theta)$ is defined in (3.15)*, with $\pi^D = 1 - \pi(a, I_2; \theta)$. $\Gamma(\bar{Y}^L) \equiv (1 - \Phi(Y^* - \bar{Y}^L))\bar{A} + \Phi(Y^* - \bar{Y}^L)M$ is the local official's reward when producing at output level $\bar{Y}^L$.*

Proof. Appendix C.1.2. □

Intuitively, the production scale of local officials' ability with low ability is small and never reaches $\bar{Y}^L$. Thus, households will not protest due to overproduction (and bad environment). For local officials, both avoiding household protest due to low consumption and increasing the probability of being promoted require higher production level. Therefore, local officials invest all their resources into production without caring about environmental quality.

After output level reaches the threshold, $\bar{Y}^L$, extra production will lead to household protest in a bad state (see Lemma 3.1). According to the cadre system, the responsible local official is demoted. This happens with probability, $1 - \pi(a, I_2; \theta) > 0$. Therefore, on the one hand, increasing production beyond the protest threshold, $\bar{Y}^L$, will result in a loss of $\Sigma^L \equiv [1 - \pi(a, I_2; \theta)] \left(\Gamma(\bar{Y}^L) - \underline{A}\right)$ for sure, where $\Gamma(\bar{Y}^L) - \underline{A}$ is the decrease in income due to demotion. On the other hand, the benefit of increasing production above $\bar{Y}^L$ is a higher probability of promotion and the corresponding income. Specifically, the probability of promotion at output level Y is given by $1 - \Phi(Y^* - Y)$. Therefore, the

increase in probability of promotion compared with producing at $\bar{Y}^L$ is $\Delta^P = \Phi(Y - \bar{Y}^L)$.[15] As Y increases from $\bar{Y}^L$, Δ^P starts to increase continuously from zero. Therefore, local officials need a minimal level of increase in promotion probability to compensate for the loss of being demoted, Σ^L. This requires a minimal increase in production. However, local officials' investment in production is bounded from above by the maximal feasible investment and also by the investment, $I_1(\bar{Y}^H, a, \theta)$ required to guarantee $\underline{U}$ and avoid protest. This means that in general it is not guaranteed that there will be local officials who produce above $\bar{Y}^L$. However, if there is a local official with ability a who finds it beneficial to produce above $\bar{Y}^L$, all officials with ability above a would also produce at higher level. In the sense, there exists a threshold of ability, $\bar{a}$, such that officials with ability higher than the threshold will choose a higher level of production than $\bar{Y}^L$. This is the case in the quantitative results illustrated in Figure 3.5.

3.4.2 The central government's equilibrium environmental policy

Similar to the social optimum benchmark, the central government knows the ability distribution of local officials and anticipates their investment behavior. The only difference is that the career-concerned officials may produce at $Y > Y^L_{sus}$, which results in negative infinite household utility if a bad state realizes. Notice that this is never the case under social optimum where local officials are benevolent and a negative infinite household utility is never optimal. Therefore, the central government's maximization problem defined in (3.12) can be reformulated as one with an additional constraint on the set of environmental policy the central government may choose from. Specifically, and the central government' choice of environmental policy has a direct impact on local officials' behavior. Specifically, to increase their probability of promotion, local officials may produce at the upper boundary for household protest, $\bar{Y}^H$. However, the central government needs to make sure that the output of local officials is never above Y^L_{sus}. According to Proposition 3.1, local officials never produce more than the upper threshold of production, $\bar{Y}^H$. Therefore, the central government only need to guarantee that $\bar{Y}^H(\theta) \leq Y^L_{sus}$. Since $\bar{Y}^H(\theta)$ increases in the policy, θ, whereas Y^L_{sus} is independent of the policy, this defines an upper bound $\bar{\theta}$: $\bar{Y}^H(\bar{\theta}) = Y^L_{sus}$.

Therefore, the central government's maximization problem in the equilibrium under

[15]Specifically, the increase in probability of promotion is $\Delta^P = 1 - \Phi(Y^* - Y) - 1 + \Phi(Y^* - \bar{Y}^L) = \Phi(Y - \bar{Y}^L)$.

the cadre system is given by:

$$\max_{\theta} \quad \int_N \mathbb{E}U(a,\theta)d\Psi(a), \tag{3.17}$$

$$\text{s.t.} \quad \lambda(\theta) \leq \bar{\lambda} \text{ and } 0 \leq \theta \leq \bar{\theta}. \tag{3.18}$$

3.4.3 Equilibrium characterization

The equilibrium in the economy is defined as follows:

Definition 3.4. *An equilibrium is characterized by the following optimal solution:*

1) Households make their protest decision according to Lemma 3.1.

2) Local officials' determine their optimal investment allocation according to Proposition 3.1.

3) The central government choose the optimal environmental policy according to (3.17)*.*

3.5 Numerical Exercises

In this section, we calculate the optimal environmental policy of the central government, the optimal investment allocation of local officials and household welfare numerically for the social optimum benchmark discussed in Section 3.3 and in the equilibrium under the cadre system as characterized in Section 3.4. We assume specific functional forms and parameter values to calibrate the model. The calibration is not meant to be a quantitative description of Chinese reality. It is indeed to be taken as an exercise for illustrating important mechanisms of environmental policy outcomes in a model that qualitatively captures the economic and institutional situation in China. By comparing the quantitative results under the two cases, we address three key questions of the paper: First, how the cadre system and local households' protest behavior influence the optimal investment allocation of local officials with different abilities. Second, how local officials' investment behavior and household protest influence the central government's choice of environmental policy. And lastly, how the environmental policy and local officials' behavior determine household welfare and environmental quality of different regions.

Households' utility is assumed to take the Cobb-Douglas form on $E \geq 0$, $U(C,E) = C^{\alpha}E^{1-\alpha}$, and is equal to negative infinity if $E < 0$. The output function exhibits a similar form, $Y(a,I_1) = Pa^{\beta}I_1^{1-\beta}$, where P is a scale parameter.

The pollution abatement technology is a linear transformation of an exponential function,

$$\varphi(\theta) = \bar{\varphi}(1 - e^{-\chi\theta}), \ \theta \in [0,1].$$

Parameter	Description	Value
α	Household's utility	0.5
β	Production parameters	0.2
P		5
$\bar{E}$	Environmental quality	3
$\bar{\varphi}$	Marginal pollution	0.5
χ	Abatement technology	2
a_{min}	Minimal ability	0.6
a_{max}	Maximal ability	6
ρ	Policy complexity coefficient	0.8
τ	Tax rate	0.2
$\underline{U}$	Protest threshold	1.8
Y^*	Promotion threshold	3.87
$\bar{A}$	Promotion reward	25
M	Incumbent reward	5
$\underline{A}$	Demotion income	3

Table 3.1: Parameter values

where $\chi > 1$ is a scale parameter. $e^{-\chi\theta} \in [0,1]$ implies that pollution reduction is a share, $1 - e^{-\chi\theta}$, of the marginal pollution without abatement, $\bar{\varphi}$. The technology satisfies the conditions in (3.1).

Lastly, the probability of successful policy implementation is given by

$$\pi(a, I_2; \theta) = a\left(1 - \rho\theta^{I_2}\right)/a_{max},\ \theta \in [0,1],$$

where $\rho \in (0,1)$, so that even extreme environmental policy (i.e., $\theta = 1$) can be implemented successfully with positive probability $a(1-\rho)/a_{max}$.[16] The expression is divided by a_{max} to ensure $\pi(a, I_2; \theta) \in [0,1]$. In addition, the functional form satisfies the conditions in Assumption 3.1, particularly that ability and investment are complementary in increasing the probability of policy success.

We specify the parameter values in Table 3.1. In particular, we choose the environmental quality, $\bar{E}$, and the marginal pollution, $\bar{\varphi}$, such that the maximal output level that allows for a sustainable environment (i.e., $E = \bar{E} - \bar{\varphi}Y \geq 0$ and thus $Y_{max} = \frac{\bar{E}}{\bar{\varphi}}$) is roughly comparable in magnitude with the attainable output by local officials.[17] A large ratio $\frac{\bar{E}}{\bar{\varphi}}$ would imply that reducing production-caused pollution has little impact on the environmental quality of the economy and thus bias local officials' investment towards

[16] If a policy can never be successfully implemented (i.e., $\exists\tilde{\theta}$, s.t. $\forall a \in [a_{min}, a_{max}]$, $\pi(a, I_2; \tilde{\theta}) = 0$), it makes no sense to choose such a policy anyhow.

[17] To see this more clearly: Using the Cobb-Douglas production function, the maximal achievable production level is given by $Y^{max}(a) = P(\tau P)^{\frac{1-\beta}{\beta}}a$. With the parameter values in Table 3.1, $\frac{\bar{E}}{\bar{\varphi}} = 6$, and $Y^{max}(a) \in [3, 30]$ for $a \in [0.6, 6]$.

production.[18] In contrast, a small $\frac{\bar{E}}{\bar{\varphi}}$ would restrict local output significantly. $\underline{U}$ is chosen such that case iii) in Lemma 3.1 is relevant. $\chi = 2$ implies that by implementing an extreme abatement technology (i.e., $\theta = 1$), 86.47% of pollution per unit of output can be reduced. The threshold of promotion, Y^*, is of similar scale to the production level that maximizes household utility in a good state.

The rest of the parameters were chosen such that numerical solutions do not run into economically unmeaningful areas like negative values. Obviously, Table 3.1 does not claim to be a calibration that matches reality in China in a quantitative sense. The purpose of the calibration is rather to allow for a numerical illustration of the mechanisms at work in the model.

3.5.1 Numerical solution in the social optimum benchmark

In this section, we characterize numerical solution of the socially optimal environmental policy and investment allocation in the social optimum benchmark. First, we solve for the optimal investment allocation of local officials, I_1 and I_2, for any given environmental policy, θ (Section 3.5.1.1). From this, we get environmental quality and household welfare for all $\theta \in (0, 1]$ (Section 3.5.1.2). And in the end, the central government calculates for each possible θ the share of household protest (according to (3.10)) and chooses the environmental policy, θ, such that there is no political instability and that aggregate households' expected utility is maximized (Section 3.5.1.3).

3.5.1.1 Socially optimal allocation of investments

In this case the local officials' behavior is characterized by the solution of (3.9). Figure 3.2 depicts for given environmental policy, θ, the optimal investment allocation, $\{I_1^*(a,\theta), I_2^*(a,\theta)\}$, of local officials with different abilities, a.[19] As ability rises, local officials allocate a larger share of investments into environment-related infrastructure (i.e., I_1^* increases and I_2^* decreases). Intuitively, for given investment in production, I_1, local officials with higher ability can generate more output. This drives down households' marginal utility of consumption. However, higher output deteriorates environment and thus drives up the marginal utility of environment. By allocating more investment into

[18] In this case, a corner solution with allocation of all resources to production-related investment may be optimal for local officials of all ability. One could consider a situation of low economic development, when output is low and environment quality is high. It benefits both households and local officials to focus on production. However, in this paper we want to characterize heterogeneous behavior of local officials, in particular, overproduction of local officials with high abilities that induces environment deterioration. Therefore, the corner solution is of less interest.

[19] For Figure 3.2 and Figure 3.3, we set as an example $\theta = 0.1274$ to be consistent with the illustration in the equilibrium under the cadre system (see footnote 22 for more details). The shape of the graph maintains for $\theta \in (0, 1]$.

environment-related infrastructure and less in production, local officials with high abilities can alleviate the negative impact of high output on environment, and increase the probability of successful policy implementation, and thus achieve maximal household utility.

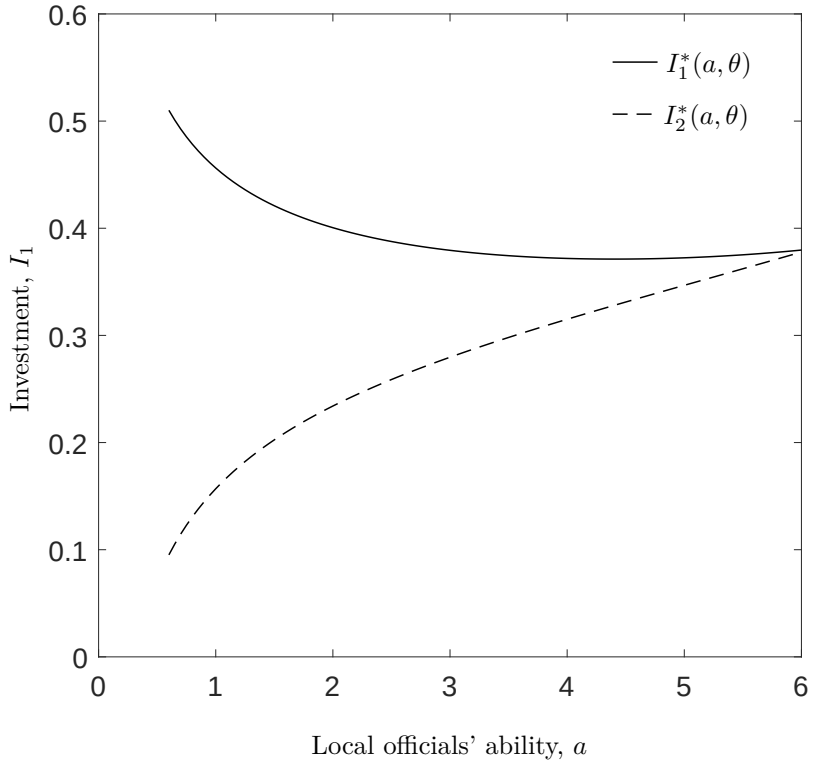

Figure 3.2: Socially optimal allocation of investments

3.5.1.2 Socially optimal environmental quality and household welfare

With the investment allocation calculated in Section 3.5.1.1, we compute household consumption, $C = (1-\tau)Y(a, I_1^*(a))$, and environmental quality in high and low states, $E^s = \bar{E} - (\bar{\varphi} - \mathbb{1}_{\{s=H\}}\varphi(\theta))Y(a, I_1^*(a)), s \in \{H, L\}$. From this, we get household utility in the two states, $U^s(C, E) = U(C, E^s), s \in \{H, L\}$.

The results are plotted in Figure 3.3. Even though local officials' investment in production-related infrastructure decreases as ability increases (see Section 3.5.1.1), the resulting output and thus household consumption is higher in regions with more able officials. This is due to a strong positive impact of high ability on local output. As a result of the high output, the environmental quality in both states is worse in regions with high-ability local officials. However, the expected environmental quality, $\mathbb{E}E(a,\theta) \equiv \pi(a, I_2; \theta)E^H + (1 - \pi(a, I_2; \theta))E^L$, is only slightly worse off, as is indicated by the downward-sloping, but much flatter curve in subplot 2. The combination of higher

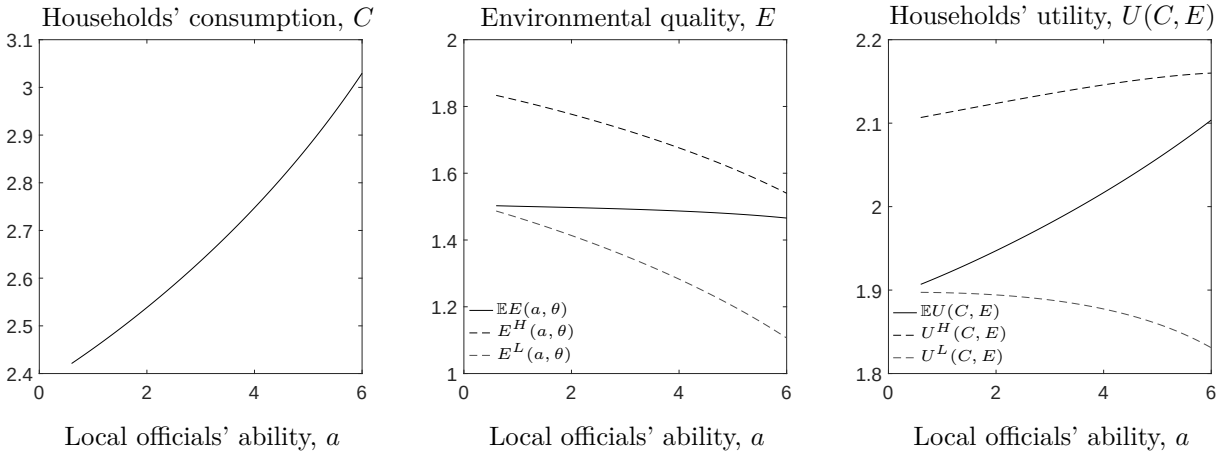

Figure 3.3: Socially optimal environmental quality and household welfare

investment in environment, $I_2^*(a,\theta)$, and higher ability increase the probability of policy success. Consequently, the gap between the expected environment quality, $\mathbb{E}E(a,\theta)$, and the environmental quality in high state, $E^H(C,E)$, shrinks when the local officials' ability is higher.

The last subplot illustrates household utility, $U^s(C,E)$, in high and low states, $s \in \{H,L\}$, and the households' expected utility, $\mathbb{E}U(C,E)$, as a function of local officials' ability, a (for a given environmental policy of the central government, θ). Not surprisingly, the household utility in a high state increases with their local official's ability, which implies that the utility gain from higher consumption outweighs the utility loss from lower environmental quality. In the low state, however, the situation is the opposite.[20] Like the expected environmental quality, the expected household utility approaches household utility in high state, because as local officials' ability increases, the probability of high state rises.

3.5.1.3 Socially optimal environmental policy

So far we solved numerically the program ((3.9) in Section 3.3.1) for the optimal investment allocation, $\{I_1^*(a,\theta), I_2^*(a,\theta)\}$, at all combinations of local officials' ability, $a \in$

[20] Essentially, we discuss the partial derivative of household utility with respect to their local official's ability. Namely, $\frac{\partial U^s(C,E)}{\partial a} = \frac{\partial U(C,E^s)}{\partial C}\frac{\partial C}{\partial a} + \frac{\partial U(C,E^s)}{\partial E}\frac{\partial E^s}{\partial a} = \left((1-\tau)\frac{\partial U(C,E^s)}{\partial C} - (\bar{\varphi} - \mathbb{1}_{\{s=H\}}\varphi(\theta))\frac{\partial U(C,E^s)}{\partial E}\right)\frac{\partial Y}{\partial a}$, where Y is the output under the socially optimal investment allocation. Since $\frac{\partial C}{\partial a} = (1-\tau)\frac{\partial Y}{\partial a} > 0$ from subplot 1, $\frac{\partial U^H(C,E)}{\partial a} > 0$ implies that $(1-\tau)\frac{\partial U(C,E^H)}{\partial C} > (\bar{\varphi} - \varphi(\theta))\frac{\partial U(C,E^H)}{\partial E}$. In other words, the marginal utility gain from higher consumption outweighs the utility loss from lower environmental quality. And $\frac{\partial U^L(C,E)}{\partial a} < 0$ implies the opposite (i.e., $(1-\tau)\frac{\partial U(C,E^L)}{\partial C} < \bar{\varphi}\frac{\partial U(C,E^L)}{\partial E}$). The change in the sign is not surprising, because as environmental quality decreases (from E^H to E^L) the marginal utility from consumption (i.e., LHS) decreases, whereas the marginal utility from environment (i.e., RHS) increases. And thus the inequality changes its direction.

$[a_{min}, a_{max}]$, and environmental policy values, $\theta \in [0, 1]$. To decide about the environmental policy, the central government needs to know the expected share of household protest, $\lambda(\theta)$, defined in (3.10) and the aggregate expected utility of households, defined in (3.11).

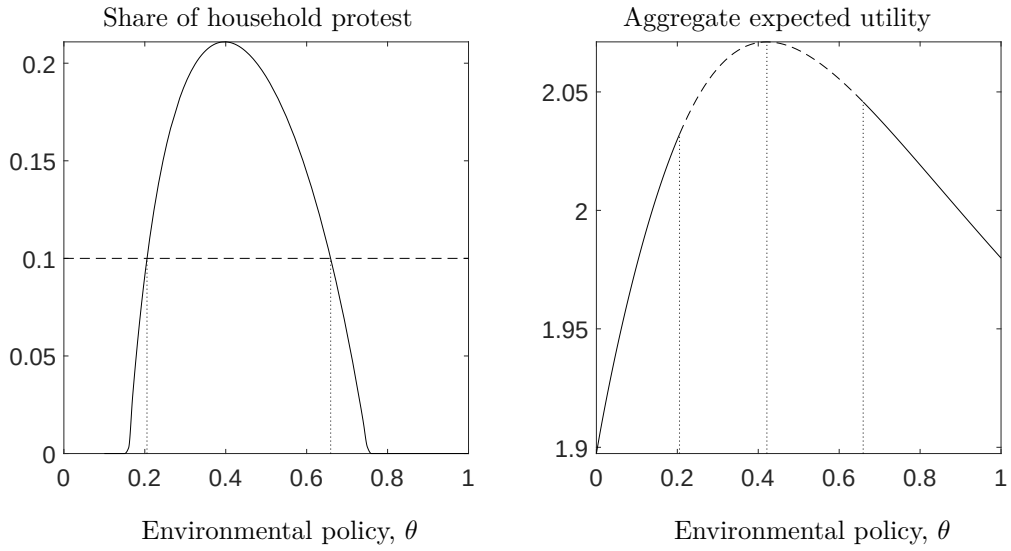

Figure 3.4: Environmental policy in the social optimum

We illustrate the quantitative results in Figure 3.4. The first subplot depicts the expected share of household protest, $\lambda(\theta)$, as a function of the environmental policy, θ. The share displays an inverse-U shape. Intuitively, when the central government chooses a moderate environmental policy (i.e., θ small), pollution abatement is low regardless of local officials' ability. Local officials take this into account when choosing the optimal investment allocation, and restrict their output level to preserve the environment. As a result household protest is relatively low at low θ. When the strength of environmental policy increases (θ rises), the abatement technology reduces pollution in a good state more significantly. Furthermore, since the task is still relatively easy, the probability that a region (especially the ones with high-ability local officials) ends up in a bad state is low. Therefore, local officials (especially high ability ones) have an incentive to overinvest in production-related infrastructure and increase household utility in high state. However, this leads to low environmental quality and thus low household utility in a bad state, which induces household protests. In the end, as the strength of environmental policy increases further, pollution abatement is very strong in a good state. However, the probability of policy success is low. In other words, with high probability the regions will end up in bad environment and thus low household utility. Anticipating this, local officials restrict their production (i.e., lower I_1^*) and shift more investment into environment-related infrastructure to increase the probability of policy success. Therefore, household utility

is less likely to fall below the threshold, and the protest share decreases.

Overall, environmental policies at the two ends restrict local officials' investment in production-related infrastructure, and reduce the share of household protest. We set the tolerable share of household protest to be $\bar{\lambda} = 10\%$; environmental policies $\theta \in (0.21, 0.66)$ induce a protest share higher than 10% and thus political instability.

Aggregate expected utility of households as a function of environmental policy, θ, is plotted in the subplot 2. Due to the negative externalities of political instability, aggregate expected utility in the interval $(0.21, 0.66)$ is zero. The dashed line represents the level of utility which would result without any negative externality of political instability. The shape of the curve is determined by the tradeoff between a better environmental quality versus a lower probability of policy success. In our calibrated model, the central government's optimal environmental policy is $\theta^* = 0.66$, at the boundary of household protest. However, as can be seen directly from the figure, without political instability consideration, the optimal environmental policy would be $\theta^{**} = 0.42$.[21]

3.5.2 Solution of equilibrium under cadre system

3.5.2.1 Equilibrium allocation of investments

Under the cadre system, local officials allocate their investment for a given environmental policy, θ, according to Proposition 3.1. The solid black line in Figure 3.5 gives the optimal investment allocation of opportunistic local officials, and the dashed lines define the boundaries of the region in which the solution must lie.[22] According to Proposition 3.1, these boundaries include the maximal feasible investment, $I_1^{max}(a, \theta)$, and the required investment in production to produce, $\bar{Y}^L$ and $\bar{Y}^H$, respectively.[23]

The local officials' investment behavior follows three patterns according to their ability.[24] For officials with low ability, $a \in [a_{min}, \underline{a})$, which corresponds to case (i) in Proposition 3.1, all tax revenues are invested into production. The intuition is as discussed in

[21]Notice that our model specification results in a higher optimal environmental policy compared to the case without political stability concern (i.e., $\theta^* > \theta^{**}$). However, this should not be taken as a general feature. If the functional forms of pollution abatement technology and probability of policy success change, the opposite may emerge (i.e., $\theta^* < \theta^{**}$).

[22]Again we use $\theta = 0.1274$ in the plots. The shape holds for $\theta \in (0, 0.1274]$. In particular, $\theta = 0.1274$ is the upper bound of the strength of the environmental policy that the central government can choose from in the cadre system equilibrium without generating infinitely negative household utility.

[23]With the Cobb-Douglas production function, we can solve for these boundaries analytically. Specifically, $I_1^{max} = (\tau P)^{\frac{1}{\beta}} a$, which displays a linear relationship with ability, a, and $I(\bar{Y}^s, a) = \left(\frac{\bar{Y}^s}{P}\right)^{\frac{1}{1-\beta}} a^{-\frac{\beta}{1-\beta}}, s \in \{H, L\}$. In addition, the threshold of ability, $\underline{a}$, is given by $\underline{a} = \frac{\bar{Y}^L}{P} (\tau P)^{-\frac{1-\beta}{\beta}}$.

[24]The two thresholds that partition the interval of $[a_{min}, a_{max}]$ into three parts are $\underline{a} = 0.79$ and $\bar{a} = 4.15$ as illustrated in Figure 3.5. In addition, $a_{min} = 0.6$ and $a_{max} = 6$.

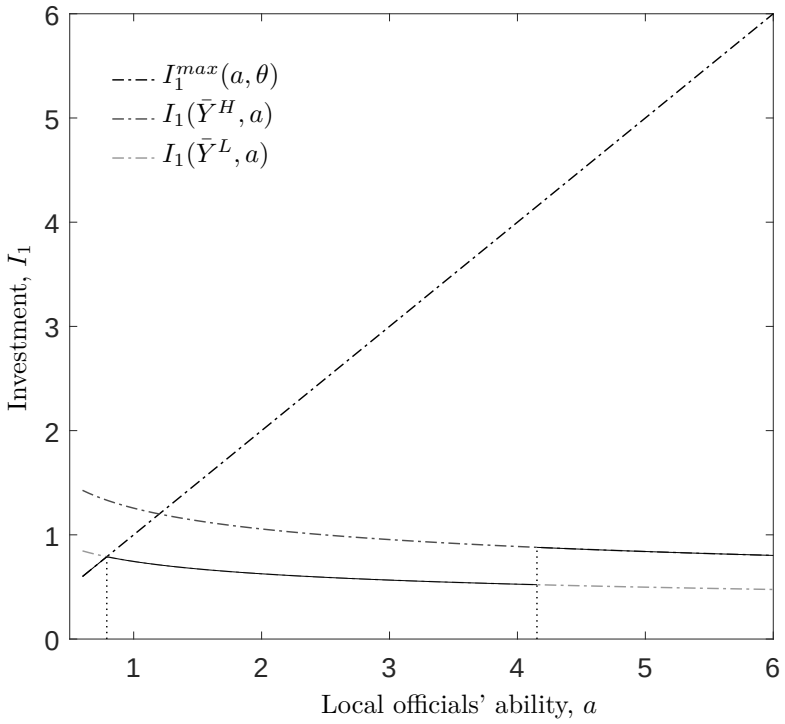

Figure 3.5: Equilibrium allocation of investments by local officials. Note: The two vertical dotted lines indicate the lower and upper threshold of local officials' ability, $\underline{a}$ and $\bar{a}$, respectively.

Section 3.4.1: The scale of production at low ability is not enough to drive household utility below the threshold that induces protest due to bad environment. In this case, local officials' motive to increase promotion probability and to avoid household protest coincide: Both encourage local officials to focus on production. Officials with medium ability $a \in [\underline{a}, \bar{a})$ produce at $\bar{Y}^L$. At this output level, households never protest regardless of state realization (see Proposition 3.1). As the local officials' ability increases within the interval, the amount of investment, $I_1(\bar{Y}^L, a, \theta)$, required to generate output, $\bar{Y}^L$, decreases, and thus the curve is downward-sloping. Furthermore, since local fiscal budget is unchanged in this interval (due to the constant production), local officials actually allocate larger share of resources into environment-related infrastructure as their ability increases. Finally, for officials with high ability, $a \in [\bar{a}, a_{max}]$, it is optimal to produce at the upper boundary of production, $\bar{Y}^H > \bar{Y}^L$. Therefore, we see a jump in the production investment, I_1. After the jump, the curve is again downward-sloping, since officials with higher abilities can generate $\bar{Y}^H$ with less resources; so they can allocate more investment into environment. Note that even though high-ability officials invest more in production than medium-ability ones, their investment in environment is not necessarily lower. Quantitative results indicate that the opposite is the case (see Figure 3.8 subplot 2 in Section 3.5.3). The reason is that for high-ability officials, one unit of extra investment

in production generates more than one unit of fiscal income. Therefore, local officials can increase investment in both production- and environment-related infrastructure.[25] The corresponding ratio of investment, I_2/I_1, by officials of the three ability categories is illustrated in Appendix C.1.3 Figure C.3.

3.5.2.2 Equilibrium environmental quality and household welfare

Having calculated local officials' investment allocation for given policies, θ, we can now calculate household consumption, regions' environmental quality in high and in low state, and household utility.

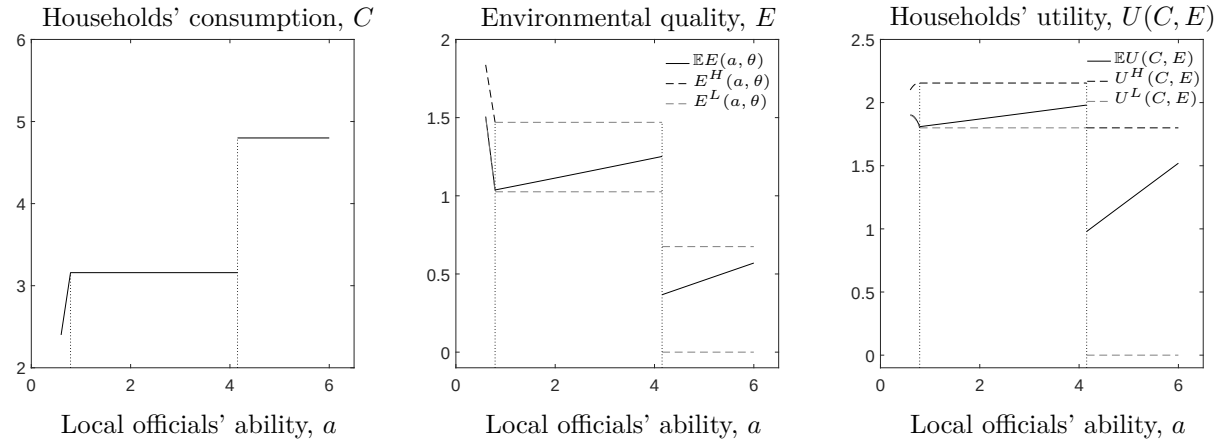

Figure 3.6: Equilibrium environmental quality and household welfare. The vertical dotted lines in the subplots indicate the lower and upper threshold of local officials' ability, $\underline{a}$ and $\bar{a}$, respectively.

Figure 3.6 illustrates the quantitative results. In regions with low-ability officials, households consume less and enjoy better environmental quality. Their utility in both states is relatively high due to high environmental quality. Notice that the households' expected utility in these regions is very close to the utility in a bad state. This mirrors the fact that the probability of a bad state is very high, because local officials invest zero budget in environment. In other words, environmental policy has very limited impact on environmental quality in low-ability regions.

Medium-ability officials produce at $\bar{Y}^L$, and thus household consumption and environmental quality in high and low states are constant, regardless of the local officials' ability within the interval. Household utility in a bad state is at the boundary of protest, $\underline{U}$. In a good state, it is above the protest threshold due to better environmental quality. Moreover, local officials within the interval allocate a larger share of investment in environment

[25] To see this more clearly: $I_2 = \tau Y(a, I_1) - I_1$, and thus the partial derivative of I_2 with respect to I_1 is given by $\frac{\partial I_2}{\partial I_1} = \tau \frac{\partial Y}{\partial I_1} - 1$. Since $\lim_{I_1 \to 0+} \frac{\partial Y}{\partial I_1} = \infty$ and $\lim_{I_1 \to \infty} \frac{\partial Y}{\partial I_1} = 0$, there exists a threshold of $\bar{I}_1$, s.t. $\tau \frac{\partial Y}{\partial I_1}|_{I_1 = \bar{I}_1} = 1$: For $I_1 < \bar{I}_1$, an increase in I_1 also increases investment I_2 and vice versa for $I_1 > \bar{I}_1$.

as their ability increases. Therefore, regions with higher-ability officials experience good environmental quality with higher probability. This explains why expected environmental quality and expected household welfare rise with ability in subplot 2 and 3.

In regions with high-ability officials, households consume $(1-\tau)\bar{Y}^H$, which is the highest level of consumption among all regions. As a consequence, there is a strong negative effect of production on environmental quality. In the chosen numerical calibration, this strong negative effect is not compensated by sufficient investment in environment so that the environmental quality is among the worst in the high-ability regions.[26] Furthermore, in these regions household utility in high state is $\underline{U}$, and in low state, the utility is lower due to bad environmental quality. This is a direct consequence of the reward scheme. Since in the case of demotion, local officials' punishment is independent of their behavior (and of household welfare), they do not care what exactly happens in their region.[27] Therefore, low utility, even environmental disaster (i.e., unsustainable environment due to overproduction, $Y > \frac{\bar{E}}{\bar{\varphi}}$ as defined in Section 3.2.3.1) could occur. This is never the case in the social optimum benchmark where local officials balance household utility in high and low states. As a result, the central government has to account for an extra constraint in the cadre system: The environmental policy must avoid to induce local production beyond the level consistent with a sustainable environment. This constraint defines a maximal environmental policy $\bar{\theta}$ that is feasible for the central government.

3.5.2.3 Equilibrium environmental policy

Similar to the social optimum case, we plot the share of household protest and aggregate expected utility of households as a function of environmental policy, θ. Furthermore, according to the three patterns of investment allocation of local officials, we decompose households into three categories: Households from high-, medium- and low-ability regions, respectively.

Notice that in equilibrium, a change in the policy not only influences the share of protest and household utility directly, but also indirectly through the change in local officials' behavior. Numerical results indicate that as the strength of environmental policy increases (i.e., θ increases), the share of local officials that produce at $\bar{Y}^H$ increases.[28]

[26]However, this is not necessarily the case in general. Even if environmental quality in high-ability regions is worse in each of the two states, L and H, respectively, expected environmental quality is not necessarily worse. This is due to higher investment in environment by high-ability officials and thus a higher probability of ending up in a good state. But one thing is for sure: Both environmental quality and household utility in a bad state are the lowest in regions with high ability officials.

[27]Note that in the high-ability regions local officials take the low risk to be demoted in a bad state in exchange for a high chance to be promoted in a good state.

[28]The share of local officials that invest all resources into production (i.e., low-ability regions) is independent of the environmental policy, and is thus unchanged.

This is the result of two counteracting effects. On the one hand, a rise in θ increases the probability of policy failure, in which case local officials that produce at high level will be demoted. This increases the expected loss, and discourages local officials from producing at high level. On the other hand, as environmental policy turns to more effective technology, local officials can produce at larger scale without inducing household protest (i.e., $\bar{Y}^H$ increases). The extra production increases local officials' probability of promotion which encourages local officials to produce at high level, and tends to decrease the threshold of the ability ($\bar{a}$) above which local officials produce at $\bar{Y}^H$. Quantitatively, the second effect dominates the first one, and thus more regions will produce at the higher level of production, $\bar{Y}^H$.[29]

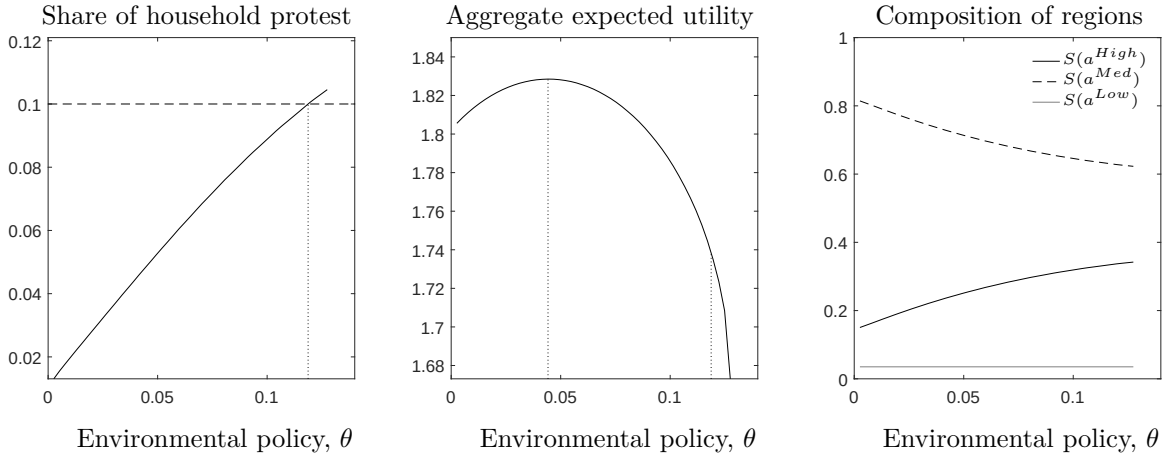

Figure 3.7: Equilibrium environmental policy. With slight abuse of notations, we use a^{Low} if $a \in [a_{min}, \underline{a})$, a^{Med} if $a \in [\underline{a}, \bar{a})$, and a^{High} if $a \in [\bar{a}, a_{max}]$.

Not surprisingly, the share of household protest increases if the central government chooses a stronger environmental policy (i.e., θ increases). On the one hand, as θ rises the probability of a bad state and thus policy failure increases. Therefore, ceteris paribus the expected share of household protest increases. On the other hand, the share of regions that produce at high level of production increases. Since in regions which produce at high level, households protest in a bad state, a larger share of household protest.

Aggregate household utility displays an inverse-U shape similar to the social optimum case. But the central government chooses from a much smaller set of environmental policy, in order to restrict local officials' incentive to produce at output levels unsustainable by the environment in a bad state (i.e., $Y > Y^L_{sus}(\theta)$). Compared with the social optimum

[29]Notice that this result relies on the specification and parameter values of the model. For example, if the probability of policy success is very sensitive to the strength of environmental policy (i.e., the elasticity is large), an increase in θ drives down the probability significantly, local officials may actually decrease their investment in production related investment, resulting a larger share of regions producing at the lower level of output, $\bar{Y}^L$.

case, the hump shape comes not only from a tradeoff between more effective pollution abatement in a good state and a lower probability of policy success, but also from a composition effect: More effective policy increases the production threshold of protest in a good state, and thus induces more local officials to overproduce. This drives down household utility. The optimal environmental policy is $\theta = 0.04$, which is much lower than the socially optimal one ($\theta^* = 0.66$).

The composition effect is verified by the third subplot. We see clearly that as the strength of environmental policy increases, the share of local officials, $S(a^{High})$, that produce at $\bar{Y}^H$ increases; the increase share comes from a decrease in officials that produce at $\bar{Y}^L$, $S(a^{Med})$.

3.5.3 Comparison between social optimum benchmark and the equilibrium solution under the cadre system

In this section, we compare investment allocation, optimal environmental policy and household welfare under the cadre system with the social optimum.

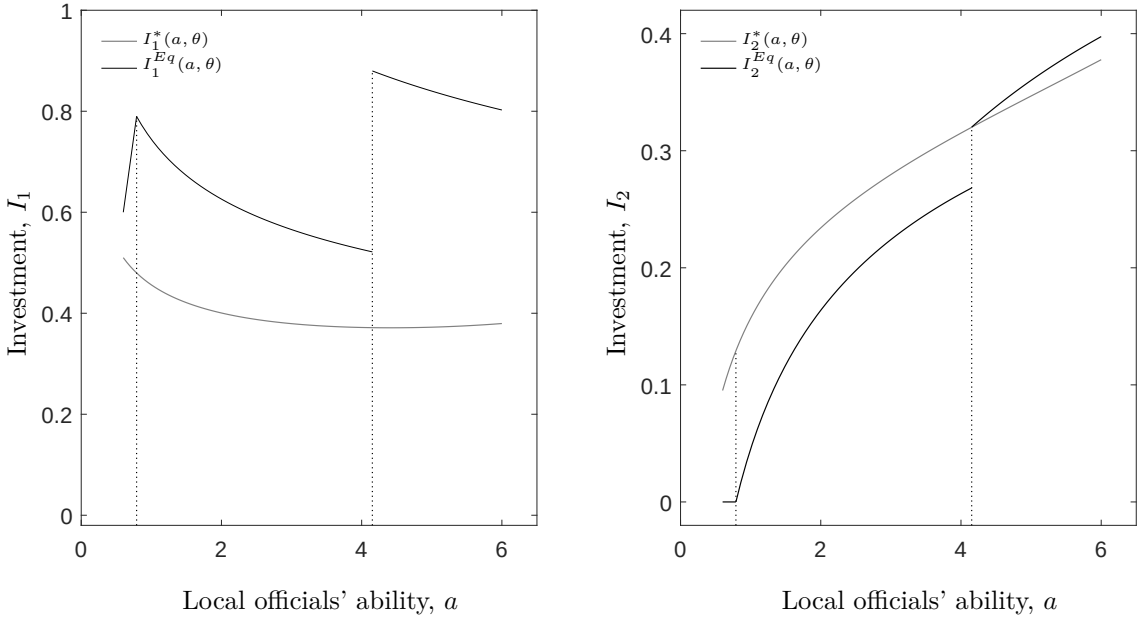

Figure 3.8: Comparison of investment allocation. The vertical dotted lines in the subplots indicate the lower and upper threshold of local officials' ability, $\underline{a}$ and $\bar{a}$, respectively.

First, we compare the local officials' investment in production and in environment for a given environmental policy. Under the cadre system, investment in production (illustrated by the solid line) is at all ability levels higher than in the social optimum (illustrated by

the dotted line) due to career motives. The effect of the career motives is especially strong with high ability officials, for whom the marginal increase in promotion probability by investing in production is the highest.

Investment in environment-related infrastructure is not uniformly lower in equilibrium under the cadre system (solid line): The investment by high ability officials is higher than under the social optimum (dotted line). However, we see from subplot 2 in Figure 3.9 that the expected environmental quality is much lower in high-ability regions. This reflects that the extra investment in environment is not enough to compensate for the extra pollution from higher production.

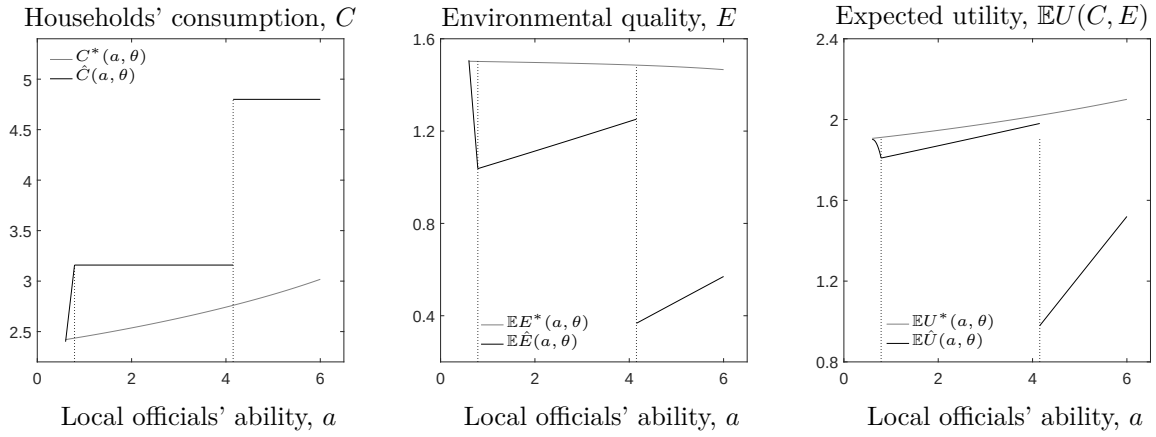

Figure 3.9: Comparison of household welfare. The vertical dotted lines in the subplots indicate the lower and upper threshold of local officials' ability, $\underline{a}$ and $\bar{a}$, respectively.

Subplot 1 in Figure 3.9 shows that households' consumption is higher in equilibrium, especially in high-ability regions as a result of local officials' overproduction. Consequently, environmental quality is lower (subplot 2). Also expected household utility is lower, as shown in subplot 3. However, one should notice that the worse environmental quality in high-ability regions is not always a result of lower investment in environment as is discussed above in Figure 3.8 subplot 2. These findings are in line with China's "high production, high pollution" situation.

Finally, if one compares the illustration of the share of household protest in Figure 3.4 and in Figure 3.7, the share is significantly lower under the social optimum, because of the benevolent behavior of local officials.[30] In addition, the optimal environmental policy is much less ambitious in equilibrium. This is mainly due to the central government's incentive to avoid overproduction of local officials. In the end, the weak environmental policy

[30]Notice that the maximal share of protest in social optimum is higher. Yet, that is under much higher level of environmental policy. If we compare the share of protest at the same level of environmental policy there is far less protest under social optimum.

together with the local officials' biased investment allocation results in a low aggregate household utility (1.83 in equilibrium, compared with 2.04 under social optimum).

3.6 Conclusion

In the last decades, China's environmental situation has attracted attention from all over the world. In a model with a central government, local officials and households, the paper analyzes the reason of China's "high output, high pollution" from a political economic perspective. The high pollution is due to distortions from two levels: Distorted allocation of tax revenues on production-related and environment-related investments by local officials, and biased environmental policy by the central government.

The model emphasizes the impact of the cadre system of the central government and of household protests on local officials' allocation of fiscal revenues on production and on environment investment. The central government promotes local officials according to local output. This induces local officials with career concerns to overproduce on the one hand. On the other hand, the possibility of household protest to some extent restricts local officials' incentive to produce and pollute. This is especially the case for medium-ability officials, for whom promotion is less likely compared with the high ability ones. However, local officials' motive to avoid household protest is inadequate to eliminate overproduction. As a result, the environment deteriorates and household welfare decreases. In the end, in order to avoid overproduction by local officials and negative externalities of household protests, the central government chooses under the cadre system an environmental policy which is less effective than the socially optimal one. This has a negative impact on environmental quality and household welfare indirectly.

Part III

Appendices

A Appendix: Chapter 1

A.1 Timing

The timing of the activities within a period T is as follows (see Figure A.1). At first, a mass of $1 - \Delta$ households is born. An endogenously determined share λ of them decides to become an entrepreneur and the rest decides to be a worker. The newborn entrepreneurs sign a lifetime binding financial contract with the banks. The banks give loans $b(V^E)$, according to the amount entitled by the respective terms of the contract, to all entrepreneurs in the economy. Entrepreneurs pay the costs of capital and labor inputs with the loans (workers consume and save by buying annuities from the banks from their labor income and capital returns) and production takes place under uncertainty. After production, entrepreneurs observe the state of their productivity realization and make a report about it to the bank. Then, entrepreneurs make state-contingent repayments $m_s(V^E)$ to the banks according to their financial contract and consume the remaining net production revenue. Further, the contract determines state-contingent promised values $V_s^E(V^E)$ as future state variable. Finally, a share $1 - \Delta$ of the workers and entrepreneurs dies and the associated firms exit.

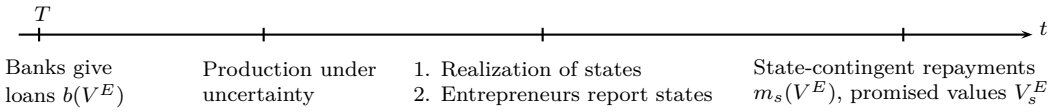

Figure A.1: Timing of terms of financial contract within one period

A.2 Derivations

A.2.1 Derivations of financial contract properties

The proofs for Proposition 1.1 and 1.2 and Lemma 1.1 and 1.2 follow the proofs on the optimal social insurance in Ljungqvist and Sargent (2000) which are based on Thomas and Worrall (1990).

A.2.1.1 Proof of Proposition 1.1

Proof. Using the definition in (1.12) and summing up $C_{s,s-1}+C_{s-1,s}$ we conclude: $C_{s,s-1}+C_{s-1,s} \geq 0$, which is equivalent to

$$U(\theta_s R(b) - m_s, L^E) - U(\theta_s R(b) - m_{s-1}, L^E) \geq U(\theta_{s-1} R(b) - m_s, L^E) - U(\theta_{s-1} R(b) - m_{s-1}, L^E) \tag{A.1}$$

Since $\theta_s > \theta_{s-1}$ and given the strict concavity of the utility function in consumption, (A.1) is satisfied only if $m_s \geq m_{s-1}$. It then follows from $C_{s,s-1} \geq 0$ that $V_s^E \geq V_{s-1}^E$. □

A.2.1.2 Proof of Lemma 1.1

Proof. Without loss of generality, we prove from the local downward constraints $C_{s,s-1} \geq 0, \forall s \in \mathcal{S}$, that for any $i > j,\ i, j \in \mathcal{S}$, $C_{i,j} \geq 0$. The case of $i < j$ can be proved from the local upward constraints $C_{s,s+1} \geq 0, \forall s \in \mathcal{S}$, using the same logic.

Proof with mathematical induction:

For $n = 1$, $C_{j+n,j} \geq 0$ holds according to the local downward constraint. Suppose for $n \geq 1$, $C_{j+n,j} \geq 0, \forall j \in \mathcal{S}$ holds; we need to prove that $C_{j+n+1,j} \geq 0$. For simplicity of notation denote $i = j + n$.

First, $C_{i,j} \geq 0$ and $C_{i+1,i} \geq 0$ are equivalent to the following inequalities:

$$U(\theta_i R(b) - m_i, L^E) + \beta \Delta V_i^E - U(\theta_i R(b) - m_j, L^E) - \beta \Delta V_j^E \geq 0,$$
$$U(\theta_{i+1} R(b) - m_{i+1}, L^E) + \beta \Delta V_{i+1}^E - U(\theta_{i+1} R(b) - m_i, L^E) - \beta \Delta V_i^E \geq 0.$$

Summing up the two inequalities we have:

$$U(\theta_{i+1} R(b) - m_{i+1}, L^E) + \beta \Delta V_{i+1}^E - \beta \Delta V_j^E + U(\theta_i R(b) - m_i, L^E) - U(\theta_i R(b) - m_j, L^E) - U(\theta_{i+1} R(b) - m_i, L^E) \geq 0. \tag{A.2}$$

Using the strict concavity of the utility function, the fact $\theta_{i+1} > \theta_i$, and $m_i \geq m_j$ from

Proposition 1.1, we have additionally the following inequality:

$$U(\theta_{i+1}R(b) - m_i, L^E) - U(\theta_{i+1}R(b) - m_j, L^E) \geq U(\theta_i R(b) - m_i, L^E) - U(\theta_i R(b) - m_j, L^E) \quad \text{(A.3)}$$

Adding (A.3) to (A.2) we have

$$U(\theta_{i+1}R(b) - m_{i+1}, L^E) + \beta\Delta V_{i+1}^E - \beta\Delta V_j^E - U(\theta_{i+1}R(b) - m_j, L^E) \geq 0.$$

Namely, $C_{i+1,j} \geq 0$. □

A.2.1.3 Proof of Lemma 1.2

Proof. First, we prove by contradiction that the local downward constraints must bind: Suppose that there exists an optimal contract $\left\{b, m_s, V_s^E\right\}_{s\in\mathcal{S}}$ such that for some $i \in \mathcal{S}$ the downward constraint does not bind (i.e., $C_{i,i-1} > 0$). Then, the general procedure is as follows: We prove that there exists a mean-preserving contraction transformation on $\left\{V_j^E\right\}_{j=i,\ldots,S}$ such that the new contract $\left\{b, m_s, \hat{V}_s^E\right\}_{s\in\mathcal{S}}$, where $\hat{V}_j^E = V_j^E$, for $j = 1, 2, \ldots, i-1$, fulfills all constraints. In particular, we make a transformation with $\sum_{s\in\mathcal{S}} \pi_s \hat{V}_s^E = \sum_{s\in\mathcal{S}} \pi_s V_s^E$, and $\hat{V}_j^E - \hat{V}_l^E \leq V_j^E - V_l^E,\ \forall j, l \in \mathcal{S}$, with at least one pair of $\{j, l\}$ giving strict inequality. In this case, under the assumption that $P(V^E)$ is strictly concave, the banks' profit increases strictly with the new contract. This contradicts the fact that $\left\{b, m_s, V_s^E\right\}_{s\in\mathcal{S}}$ is an optimal contract.

Now we describe explicitly the procedure of performing a mean-preserving contraction transformation on the contract:

Keeping $\left\{m_{i-1}, m_i, V_{i-1}^E\right\}$ as before, we decrease V_i^E until $C_{i,i-1} = 0$. Since changing V_i^E will influence the local downward incentive constraints for $s = i+1$ and sequentially $s = i+2, \ldots, S$, we decrease for each $s = i+1, \ldots, S$, V_s^E such that $C_{s,s-1} = 0$. As a result we have a new sequence of future promised value $\left\{V_s^{E\prime}\right\}_{s\in\mathcal{S}} = \left\{V_1^E, V_2^E, \ldots, V_{i-1}^E, V_i^{E\prime}, V_{i+1}^{E\ \prime}, \ldots, V_S^{E\prime}\right\}$. Now we add a positive constant, $\bar{v}$, to the sequence of future promised value, such that the promise keeping constraint is regained. Let $\hat{V}_s^E = V_s^{E\prime} + \bar{v}$. We have a new contract $\left\{b, m_s, \hat{V}_s^E\right\}_{s\in\mathcal{S}}$.

First, note that the new contract fulfills the local upward constraints automatically given the strict concavity of the utility function and the fact that $C_{s,s-1} = 0\ \forall s \in \mathcal{S}$ (see argumentation in the last part of this proof). In addition, the promise keeping constraint is still fulfilled due to the mean-preserving transformation, and the limited liability constraints are uninfluenced since b and $\{m_s\}_{s\in\mathcal{S}}$ are unchanged. Finally, for any $j = i, \ldots, S$, V_{j+1}^E must decrease at least as much as V_j^E to guarantee that $C_{j+1,j} = 0$.

Therefore, for any $j = i, \ldots, S$, $\bar{v} \leq V_j^E - V_j^{E\prime}$, indicating that $\hat{V}_j^E \leq V_j^E$ and remember that for $j = 1, 2, \ldots, i-1$ $\hat{V}_j^E = V_j^E$,. Since $\left\{V_s^E\right\}_{s\in\mathcal{S}}$ fulfills the credibility constraints, so does the new contract.

Further, notice from the procedure that the gap of the promised values between two successive states, s and $s-1$, is either unchanged or decreased, with a definite decrease in $\hat{V}_i^E - \hat{V}_{i-1}^E$. Following this we know $\forall j, l \in \mathcal{S}$, $V_j^E - V_l^E$ is non-increasing. Thus, the new contract is a mean-preserving contraction. This contradicts that $\left\{b, m_s, V_s^E\right\}_{s\in\mathcal{S}}$ is an optimal contract: We know that the local downward constraints always bind.

Given that $C_{s,s-1} = 0, \forall s \in \mathcal{S}$, rewriting the constraint we have

$$\beta\Delta(V_s^E - V_{s-1}^E) = U(\theta_s R(b) - m_{s-1}, L^E) - U(\theta_s R(b) - m_s, L^E)$$

Since $\theta_{s-1} < \theta_s, m_{s-1} \leq m_s$ and the utility function is strictly concave, we have

$$\begin{aligned} &U(\theta_{s-1}R(b) - m_{s-1}, L^E) - U(\theta_{s-1}R(b) - m_s, L^E) \geq \\ &\quad U(\theta_s R(b) - m_{s-1}, L^E) - U(\theta_s R(b) - m_s, L^E) = \beta\Delta(V_s^E - V_{s-1}^E), \end{aligned}$$

where strict inequality holds for $m_{s-1} < m_s$. Therefore, we have directly from this that the local upward constraint is never binding. Namely, $C_{s-1,s} > 0, \forall s \in \mathcal{S}$. □

A.2.1.4 Proof of Proposition 1.2

Proof. The non-decreasing entrepreneurs' utility is direct result of the binding local downward constraints.

The non-decreasing profit of banks is proved by contradiction. Suppose for the optimal contract, $\left\{b, m_s, V_s^E\right\}_{s\in\mathcal{S}}$, there exists $i, j \in \mathcal{S}, i > j$, such that

$$-b + m_i + \frac{\Delta}{1+r}P(V_i^E) < -b + m_j + \frac{\Delta}{1+r}P(V_j^E).$$

Substituting (m_i, V_i^E) with (m_j, V_j^E) increases banks' profit in state i. Since the downward constraint binds, $C_{i,j} = 0$, the terms of contract, (m_j, V_j^E), entitle the entrepreneurs the same promised value as (m_i, V_i^E). This means that we find an improvement that increases the profit of the banks without violating any constraints. This contradicts the optimality of the original contract. Therefore, in the optimal contract the banks' profits cannot decline with a higher productivity realization. □

A.2.1.5 Proof of Proposition 1.3

We use the Lagrangian. For simplification, we derive the Lagrangian and the F.O.C. for the case of two states (i.e., $\mathcal{S} = \{l, h\}$ with $\pi_h = \pi, \pi_l = 1 - \pi$). The Lagrangian is given by:

$$\begin{aligned}
\mathcal{L} = \max_{\{b, m_s, V_s^E\}_{s\in\mathcal{S}}} & -b + \sum_{s=\{l,h\}} \pi_s \left[m_s + \frac{\Delta}{1+r} P(V_s^E) \right] \\
& + \lambda_1 \left\{ \sum_{s=\{l,h\}} \pi_s [U(c_{ss}, L^E) + \beta \Delta V_s^E] - V^E \right\} \\
& + \lambda_2 \left\{ U(c_{hh}, L^E) + \beta \Delta V_h^E - U(c_{hl}, L^E) - \beta \Delta V_l^E \right\} \\
& + \lambda_3 \left\{ U(c_{ll}, L^E) + \beta \Delta V_l^E - U(c_{lh}, L^E) - \beta \Delta V_h^E \right\} \\
& + \lambda_4 c_{hh} + \lambda_5 c_{ll} \\
& + \lambda_6 (V_{max}^E - V_l^E) + \lambda_7 (V_l^E - V_{min}^E) \\
& + \lambda_8 (V_{max}^E - V_h^E) + \lambda_9 (V_h^E - V_{min}^E)
\end{aligned}$$

where $c_{ij} = \theta_i R(b) - m_j$. $\lambda_i \geq 0,\ i = 1, \ldots, 9$ are the Lagrangian multiplier for (PK), (IC), (LL) and (CC), respectively. The F.O.C.s are

$$\frac{\partial \mathcal{L}}{\partial m_h} = \pi - (\lambda_1 \pi + \lambda_2) U'(c_{hh}) + \lambda_3 U'(c_{lh}) - \lambda_4 = 0; \tag{A.4}$$

$$\frac{\partial \mathcal{L}}{\partial m_l} = (1 - \pi) - (\lambda_1 (1 - \pi) + \lambda_3) U'(c_{ll}) + \lambda_2 U'(c_{hl}) - \lambda_5 = 0; \tag{A.5}$$

$$\frac{\partial \mathcal{L}}{\partial V_h^E} = \frac{\pi}{1+r} P'(V_h^E) + (\lambda_1 \pi + \lambda_2 - \lambda_3)\beta - \lambda_8 + \lambda_9 = 0; \tag{A.6}$$

$$\frac{\partial \mathcal{L}}{\partial V_l^E} = \frac{1-\pi}{1+r} P'(V_l^E) + (\lambda_1 (1 - \pi) - \lambda_2 + \lambda_3)\beta - \lambda_6 + \lambda_7 = 0; \tag{A.7}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial b} = & -1 + \{(\lambda_1 \pi + \lambda_2) U'(c_{hh}) - \lambda_2 U'(c_{hl}) + \lambda_4\} \theta_h R'(b) \\
& + \{(\lambda_1 (1 - \pi) + \lambda_3) U'(c_{ll}) - \lambda_3 U'(c_{lh}) + \lambda_5\} \theta_l R'(b) = 0.
\end{aligned} \tag{A.8}$$

These together with the complementary conditions and $\lambda_i \geq 0,\ i = 1, \ldots, 9$ characterize the conditions an optimal contract needs to fulfill. For the following result, we consider only non-binding credibility constraints (CC) so that we have $\lambda_i = 0,\ i = 6, \ldots, 9$. Furthermore, we use that the downward constraint always binds ($\lambda_2 \geq 0$) whereas the upward constraint for $m_s > m_{s-1}$ never does ($\lambda_3 = 0$).

Proof of Proposition 1.3. From (A.4), (A.5) and (A.8) follows

$$\mathbb{E}(\theta)R'(b) + [\lambda_3 U'(c_{lh}) - \lambda_2 U'(c_{hl})](\theta_h - \theta_l)R'(b) = 1$$

Since $\lambda_3 = 0$ we have

$$\mathbb{E}(\theta)R'(b) - \lambda_2 U'(c_{hl})(\theta_h - \theta_l)R'(b) = 1. \tag{A.9}$$

Since $\lambda_2 \geq 0$ and $U'(c) \geq 0$, we have that $\mathbb{E}(\theta)R'(b) \geq 1$ and thus $b \leq b^*$, where b^* is the efficient bank loan level defined in (1.15). In addition, as $c_{hl} = \theta_h R(b) - m_l$ increases (e.g., if m_l decreases), b approaches the efficient level, b^*.

□

A.2.2 Derivation of the good market clearing condition

The goods market is cleared if aggregate output equals the sum of consumption of all households and aggregates investment (i.e., replacement of depreciated capital in stationary case). Remember from (1.29) that the formal condition is $\lambda Y = \lambda C^E + (1-\lambda)C^W + \lambda\delta K^D$. To prove that the goods market clearing condition can be derived from the other equations, we need only to prove that the RHS of (1.29) can be simplified to λY.

We aggregate the consumption of entrepreneurs C^E and workers C^W. Plugging in entrepreneur's consumption $c(.)$ from (1.7) into (1.24) (using $R(b) = R(b(V^W))$ and $m_s = m(V^E, \theta_s)$) we obtain for aggregate consumption of entrepreneurs:

$$C^E = \sum_{\tau=0}^{\infty}(1-\Delta)\Delta^\tau \sum_{s\in\mathcal{S}} \pi_s \int \left[\theta_s R(b(V^E)) - m(V^E, \theta_s)\right] d\Psi_\tau(V^E) = Y - M, \tag{A.10}$$

where the second equality follows from (1.23) and (1.22).

A cohort τ worker's consumption is $c_\tau = wl_\tau + A_\tau - p^A A_{\tau+1}$ (follows from the budget constraint (1.3)). Aggregation gives

$$\begin{aligned} C^W &= \sum_{\tau=0}^{\infty}(1-\Delta)\Delta^\tau(wl_\tau + A_\tau - p^A A_{\tau+1}) \\ &= wL^S - D + \sum_{\tau=0}^{\infty}(1-\Delta)\Delta^\tau A_\tau \\ &= wL^S - D + \Delta\sum_{\tau=0}^{\infty}(1-\Delta)\Delta^{\tau-1} A_\tau \\ &= wL^S - D + \Delta\frac{1}{p^A}D \\ &= wL^S + rD, \end{aligned} \tag{A.11}$$

where the second and forth equality follow from (1.18) and (1.17) and the annuity price $p^A = \frac{\Delta}{1+r}$ from (1.1) is used.

Entrepreneurs have the bank loans to finance the production costs. All money is used for production costs. Thus, the constraint in entrepreneurs' decision problem in (1.5) is binding so that we have $b(V^E) = wl^*(V^E) + (r+\delta)k^*(V^E)$. In the aggregate, this means

$$B = wL^D + (r+\delta)K^D. \tag{A.12}$$

Plugging (A.10) and (A.11) into the RHS of (1.29) gives:

$$\begin{aligned}
\text{RHS} &= \lambda(Y - M) + (1-\lambda)(wL^S + rD) + \lambda\delta K^D \\
&= \lambda Y - \lambda M + (1-\lambda)wL^S + (1-\lambda)rD + \lambda rE - \lambda rE + \lambda\delta K^D \\
&= \lambda Y + \lambda wL^D + \lambda(r+\delta)K^D - \lambda(M + rE) \\
&= \lambda Y + \lambda B - \lambda B \\
&= \lambda Y = \text{LHS},
\end{aligned}$$

where the equilibrium conditions (1.27) and (1.28) were used in the third equality and (1.25) and (A.12) were used in the third. This closes the proof that the goods market clearing condition can be derived from clearing in the capital and labor markets.

A.3 Numerical procedure

In this section we describe the dynamic programming algorithm for solving the partial equilibrium (i.e., workers' decision problem in Section A.3.1 and banks' optimal contract in Section A.3.2), the procedure to simulate the entrepreneurs' life path and calculate the aggregate variables (Section A.3.3), and the algorithm to calculate the unique stationary general equilibrium (Section A.3.4). All computations are done with Matlab.

A.3.1 Workers

1. Use constant r and w.

2. Set a grid for the state variable A. *Agrid* denotes the grid points of A. We set $A = [0, 10]$ and generate $nA = 50$ Chebyshev grid points on the interval.[1] We manually replace the lowest Chebyshev point with the lower bound of $A = 0$.

[1] See footnote 40 for an explanation of Chebyshev points.

3. Give an initial guess for the functional form of the value function, $V^W(A;r,w)^0$, of the policy functions $l(A)^0$ and $A'(A)^0$ and of $c(A)^0$.[2] We use $V^W(A_i)^0 = -\exp(-A_i) - 0.1$, $c(A_i)^0 = 0.1$, $l(A_i)^0 = 0.6$ and $A'(A_i)^0 = 0$ for each $A_i \in Agrid$, $i \in \{1, ..., nA\}$.

4. Solve on each grid point $A_i \in Agrid$, $i \in \{1, ..., nA\}$, the worker's problem in (1.2) subject to (1.3), $c(A_i) \geq 0$, $l(A_i) \in [0,1]$ and $A'(A_i) \geq 0$. This gives us the optimal solution of the system, $\{c(A_i)^1, l(A_i)^1, A'(A_i)^1\}$ and the corresponding updated value function $V^W(A_i;r,w)^1$ at $A_i \in Agrid$, $i \in \{1, ..., nA\}$.[3] To calculate the updated value function, we interpolate on $V^W(A;r,w)^0$ to get values for $A'(A)$ which lie between two $Agrid$-points.[4]

5. Compare the two successive iterations of value functions, $V^W(A;r,w)^1$ with $V^W(A;r,w)^0$, by defining a distance measure d_{V^W}, such that $d_{V^W} \equiv \max_{i\in\{1,...,nA\}} \left|V^W(A_i;r,w)^1 - V^W(A_i;r,w)^0\right|$. If $d_{V^W} \leq \epsilon_P$, the optimal solution from the current iteration solves the workers' problem and go to Step 5.[5] If $d_{V^W} > \epsilon_P$ go to Step 3 by updating $V^W(A;r,w)^0 = V^W(A;r,w)^1$, $c(A)^0 = c(A)^1$, $l(A)^0 = l(A)^1$ and $A'(A)^0 = A'(A)^1$ as the new starting values.

6. Save the value function $V^W(A_i;r,w) = V^W(A_i;r,w)^1$, the $A'(A_i) = A'(A_i)^1$ and $l(A_i) = l(A_i)^1$ and $c(A_i) = c(A_i)^1$ for each $A_i \in Agrid$, $i \in \{1, ..., nA\}$.

A.3.2 Financial contract

1. Use constant r and w.

2. Set a grid for the state variable V^E on the interval $\left[V^E_{min}, V^E_{max}\right]$. $V^E grid$ denotes the $nV^E = 50$ Chebyshev grid points of V^E on the interval.[6] We manually replace the lowest Chebyshev point with the lower bound V^E_{min}.

3. Make an initial guess of the functional form of the value function, $P(V^E;r,w)^0$, and of the policy functions, $\left\{b(V^E)^0, m_s(V^E)^0, V^E_s(V^E)^0\right\}_{s\in\{h,l\}}$. We use $P(V^E_i;r,w)^0 = \log(-V^E_i)$, $b(V^E_i)^0 = 1$, $m_h(V^E_i)^0 = 3$, $m_l(V^E_i)^0 = 1$, $V^E_h(V^E_i)^0 = V^E_l(V^E_i)^0 = V^E_i$ for each $V^E_i \in V^E grid$, $i \in \{1, ..., nV^E\}$.

---

[2]Even though the functions (e.g., value functions, $V^W(A)$ and $P(V^E)$) are continuous per se, they can only be evaluated on the discrete grid points in the numerical exercise. Namely, it is a mapping of each grid point into a number. This applies in all algorithms.

[3] We apply the *fmincon*-command, which finds the minimum of a constrained nonlinear multivariable function using the interior point algorithm.

[4] We use spline interpolation, which is a cubic interpolation of the values of neighbor-points.

[5]We set the tolerated distance for ending the iterations, $\epsilon_P = 0.0001$.

[6]See footnote 40 for an explanation of Chebyshev points.

4. Solve for each $V_i^E \in V^E grid$, $i \in \{1, ..., nV^E\}$ the optimal contract in (1.11) subject to (PK), (IC), (LL) and (CC).[7] This gives the optimal contract at each $V_i^E \in V^E grid$, $i \in \{1, ..., nV^E\}$, $\left\{b(V^E)^1, m_s(V^E)^1, V_s^E(V^E)^1\right\}_{s\in\{h,l\}}$, and thus the updated value function, $P(V^E; r, w)^1$.[8]

5. Compare the two successive iterations of value functions, $P(V_i^E; r, w)^1$ with $P(V_i^E; r, w)^0$, by defining a distance measure d_P, such that $d_P \equiv \max_{i\in\{1,...,nV^E\}} \left|P(V_i^E; r, w)^1 - P(V_i^E; r, w)^0\right|$. If $d_P \leq \epsilon_P$, then take the current iteration of value function and policy functions as the solution and go to Step 6. If $d_P > \epsilon_P$, start over with Step 3 by updating $P(V^E; r, w)^0 = P(V^E; r, w)^1$ and $b(V^E)^0 = b(V^E)^1$, $m_h(V^E)^0 = m_h(V^E)^1, m_l(V^E)^0 = m_l(V^E)^1$, $V_h^E(V^E)^0 = V_h^E(V^E)^1$ and $V_l^E(V^E)^0 = V_l^E(V^E)^1$ as the new starting value for the next iteration.

6. Save the value function $P(V^E; r, w) = P(V^E; r, w)^1$ and the optimal contract $b(V^E) = b(V^E)^1$, $m_h(V^E) = m_h(V^E)^1, m_l(V^E) = m_l(V^E)^1$, $V_h^E(V^E) = V_h^E(V^E)^1$ and $V_l^E(V^E) = V_l^E(V^E)^1$.

A.3.3 Life path simulation and equilibrium variables

In this appendix, we simulate entrepreneurs' life paths and calculate the aggregate variables related to entrepreneurs by combining the optimal solution from Section A.3.2 and the entrepreneurs' decision.[9] Moreover, we calculate workers' aggregate deposits D and the labor supply L^S according to the straightforward analytical expressions (1.17) and (1.18).[10] Using the aggregate variables, we calculate equilibrium values for the entrepreneurial share.

1. Simulate for $N^E = 10,000,000$ entrepreneurs' age and life paths with history of productivity realizations. We use two random numbers, u_t^i and o_t^i, to denote en-

[7] For, (IC) we put in the constraint only the binding local downward constraint, since by the result of Lemma 1.2 the local upward constraint is never binding for the optimal contract.

[8] As in the algorithm for solving the workers' problem, we apply the *fmincon*-command to solve for the optimal contract at each grid point and use spline interpolation to calculate the value function for the next iteration.

[9] The numerical procedure described in this section is performed for given interest rate and wage rate, $\{r, w\}$.

[10] Calculating the aggregate deposits and the aggregate labor supply according to analytical expressions instead of applying life path simulation is simply to save computational time. Notice that to guarantee that the equilibrium factor prices approximate the true values under acceptable computational error, ϵ_{GE}, we only need to make sure that the computed aggregate values approximate the true values at higher precision, irrespective of the way they are calculated. And this is guaranteed in Step 4 for the aggregate deposits and the aggregate labor supply (i.e., $\epsilon_L = 0.0001 < \epsilon_{GE}$). For the aggregate capital and labor demand, on the other hand, we have checked that by increasing N^E to ten times of the current number, the changes of these two aggregate values are below $\epsilon_L = 0.0001$ as well.

trepreneur i's productivity realization and death / survival at time t, respectively. The procedure is as follows:

(a) Start from entrepreneur $i = 1$, period $t = 1$.

(b) Use a random number generator to generate two numbers o_1^i and u_1^i, which are uniformly distributed on the interval [0, 1].

(c) If $u_1^i < \pi_l$, save the entrepreneur's productivity realization in this period as low, $\theta_t^i = \theta_l$, and otherwise as high $\theta_t^i = \theta_h$.[11]

(d) If $o_t^i < \Delta$, the entrepreneur survives to the next period, increase t by one and go back to Step 1b. If $o_t^i > \Delta$, the life path stops. Save her year of life, $A^i = t$, go to Step 1b and simulate for the next entrepreneur $i + 1$.

(e) Save all entrepreneurs' years of life, $\{A^i\}_{i=1}^{N^E}$, and the sequence of productivity realizations, $\{\theta_t^i\}_{t=1}^{A^i}, i = 1, \ldots, N^E$.

2. Using the simulation of the life paths from Step 1 and the policy function from Section A.3.2, we determine the corresponding promised value V^i, bank loans $b(V^i)$ and the repayments $m(V^i, \theta_{A^i}^i)$ for each entrepreneur i in their last period in life $t = A^i$. Notice that the promised utility relevant for calculating the bank loans and repayments is the value at the beginning of the last period. This means that the productivity realization in $t = A^i$, $\theta_{A^i}^i$, is only used for calculating the repayments (see Step 2d). Specifically, the procedure is as follows:

(a) Set $V_0^i = V^W(0; r, w)$ using the workers' value function from Section A.3.1.

(b) Start from entrepreneur i, period $t = 1$.

(c) If $t \leq A^i - 1$, the promised utility of entrepreneur i at the beginning of period t is $V_t^i = V_s^E(V_{t-1}^i)$, where $V_s^E(V)$ is entrepreneurs' transition function solved in Section A.3.2. Repeat the step until the condition is no longer satisfied.

(d) Calculate the optimal banks loans and repayments, $\{b^i, m^i\}_{i=1}^{N^E}$, using the policy functions solved in Section A.3.2, the productivity realization in the last period of life, $\{\theta_{A^i}^i\}_{i=1}^{N^E}$ from Step 1, and the promised utility from Step 2c, $\{V_{A^i-1}^i\}_{i=1}^{N^E}$. Specifically, $b^i = b(V_{A^i-1}^i)$ and $m^i = (V_{A^i-1}^i, \theta_{A^i}^i)$.

3. Calculate the aggregate bank loans B and repayments M. Specifically, we aggregate banks loans b^i and repayments m^i over all entrepreneurs $i = 1, \ldots, N^E$ and divide the two sums by N^E to normalize the mass of the population to 1. Further, we calculate the aggregate capital demand K^D and aggregate labor demand L^D according to (1.33).

[11] We set π_l=1/2 and $\Delta = 0.92$.

4. Calculate workers' aggregate deposits D and the labor supply L^S according to equation (1.17), (1.18) and the optimal decisions solved in A.3.1. Specifically, start from the deposits and labor supply of workers of age $t = 1$, $Sum_A = p_A A'(A_{t-1})$ and $Sum_L = l(A_{t-1})$, weighted by their population size, $(1-\Delta)\Delta^{t-1}$. Constantly add to Sum_A and Sum_L the weighted deposits and labor supply of the older generation, until the differences between the sums in two successive iterations is below $\epsilon_L = 0.0001$, respectively.[12]

5. Determine the share of entrepreneurs from the labor market condition, $\lambda = \frac{L^D}{L^D+L^S}$, the zero-profit condition, $P(V^w(0;r,w);r,w)$, and the excess capital demand, $X(r,w) \equiv \lambda K^D - (1-\lambda)D - \lambda\frac{B-M}{r}$. Notice that all endogenous variables, $\left\{\lambda, K^D, D, B, M\right\}$ are functions of the factor prices, (r,w).

A.3.4 Numerical procedure for general equilibrium

In this section, we characterize the procedure for solving the general equilibrium (i.e., the factor prices (r,w)). The theoretical ground and intuition of the algorithm are given in section A.4. We will mention banks' profit at $V^E = V^W(0;r,w)$, $P(V^W(0;r,w);r,w)$, repeatedly in this section. To save notation we write $\Pi(r,w)$ (or Π when the specific values of (r,w) are irrelevant) instead of the full expression.

1. Start with (r_0, w_0) as an initial guess for the equilibrium factor prices. Calculate the partial derivatives of the banks' profit, Π, and of the excess capital demand, X, at (r_0, w_0). Denote the values of the partial derivatives as S_r^Π, S_w^Π, S_r^X, and S_w^X, respectively.[13] Set $w_L = w_U = NaN$, where NaN represents undefined numerical results in Matlab.

2. Set $r_L = r_U = NaN$.

3. Set $(r', w') = (r_0, w_0)$. Calculate $V^W(0;r',w')$ according to Algorithm A.3.1, and $P(V^E;r',w')$ according to Algorithm A.3.2.

[12] We set $\epsilon_L << \epsilon_{GE}$ so that the equilibrium is not susceptible to calculation error in the workers' aggregate variables.

[13] To approximate the partial derivatives, we calculate the value of $\Pi(r,w)$ and $X(r,w)$ at (r_0,w_0), $(r_0+\varepsilon, w_0)$ and $(r_0, w_0+\varepsilon)$, respectively. Applying the definition of partial derivatives, we have $S_r^\Pi \approx \frac{\Pi(r_0+\varepsilon,w_0)-\Pi(r_0,w_0)}{\varepsilon}$, $S_w^\Pi \approx \frac{\Pi(r_0,w_0+\varepsilon)-\Pi(r_0,w_0)}{\varepsilon}$, $S_r^X \approx \frac{X(r_0,w_0+\varepsilon)-X(r_0,w_0)}{\varepsilon}$, and $S_w^X \approx \frac{X(r_0,w_0+\varepsilon)-X(r_0,w_0)}{\varepsilon}$. Notice that this procedure is time consuming, because we need to calculate through the entire model at each combination of (r,w). Since the searching region for the general equilibrium is relatively small (within interval of magnitude 0.01), the change in partial derivatives is small. Therefore, we use these values as an approximation of the partial derivatives in all iterations to save computational time.

4. If $|\Pi(r', w')| < \epsilon_{GE}$, the banks' profit is close enough to zero and go to Step 6.[14] Otherwise, if $\Pi(r', w') > 0$, save $r_L = r'$ and $\Pi_L = \Pi(r', w')$. If $\Pi(r', w') < 0$, save $r_U = r'$ and $\Pi_U = \Pi(r', w')$.[15]

5. If neither r_L nor r_U is NaN, let $r_0 \equiv \frac{r_U \Pi_L - r_L \Pi_U}{\Pi_L - \Pi_U}$. Otherwise, $r_0 \equiv r' - \Pi(r', w')/S_r^P$. Go to Step 3.

6. Calculate the excess capital demand $X(r', w')$ and share of entrepreneurs $\lambda(r', w')$ according to Section A.3.3.

7. If $|X(r', w') - \Pi(r', w')| < \epsilon_{GE}$, then the (r', w') and the corresponding $\lambda(r', w')$ in the current iteration are the equilibrium.[16] Otherwise, if $X(r', w') < 0$, save $w_L = w'$ and $X_L = X(r', w')$. If $X(r', w') > 0$, save $w_U = w'$ and $X_U = X(r', w')$. [17]

8. If neither w_L nor w_U is NaN, let $w_0 \equiv (w_L + w_U)/2$, and $r_0 \equiv r' - \left(\Pi(r', w') + S_w^P(w_0 - w')\right)/S_r^P$. Otherwise, we set

$$\begin{pmatrix} r_0 \\ w_0 \end{pmatrix} \equiv \begin{pmatrix} r' \\ w' \end{pmatrix} - A^{-1}b, \tag{A.13}$$

where $A = \begin{pmatrix} S_r^P & S_w^P \\ S_r^X & S_w^X \end{pmatrix}$, and $b = \begin{pmatrix} \Pi(r', w') \\ X(r', w') \end{pmatrix}$.[18] Go to Step 2.

A.4 Theoretical ground and intuition of Algorithm A.3.4

The theoretical ground of the stationary general equilibrium searching in Algorithm A.3.4 is based on the continuity of the aggregate variables and the values functions with respect

[14] We set $\epsilon_{GE} = 0.001$.

[15] An intuitive description of this step is given in Section A.4.

[16] Notice that by setting the criterion as $|X(r, w) - \Pi(r, w)| < \epsilon_{GE}$ instead of $|X(r, w)| < \epsilon_{GE}$, we decreases the computational error of the equilibrium factor prices. Essentially, we want to avoid the case when both $X(r, w)$ and $\Pi(r, w)$ are marginally below ϵ_{GE} but of the opposite sign. By analysis similar as illustrated in Figure A.2, this deviates the numerical solution from the true values much more than if both $X(r, w)$ and $\Pi(r, w)$ are marginally below ϵ_{GE} but of the same sign.

[17] An intuitive explanation of this step is given in footnote 22 and the corresponding part in the main text.

[18] We apply Tayor's expansion on $\Pi(r, w)$ and $X(r, w)$. Namely, $\Pi(r_0, w_0) \approx \Pi(r', w') + \frac{\partial \Pi}{\partial r}(r_0 - r') + \frac{\partial \Pi}{\partial w}(w_0 - w')$, and $X(r_0, w_0) \approx X(r', w') + \frac{\partial X}{\partial r}(r_0 - r') + \frac{\partial X}{\partial w}(w_0 - w')$. Setting $\Pi(r_0, w_0)$, $\Pi(r', w')$, and $X(r_0, w_0)$ to be zero we get equation (A.13).

to (r, w), and numerical properties of the zero-profit condition, $\Pi \equiv P(V^W(0; r, w); r, w)$, and the excess capital demand, $X \equiv \lambda(r, w)K^D(r, w) - (1-\lambda(r, w))D(r, w) - \lambda(r, w)E(r, w)$.[19]

Property 2. *The zero-profit condition and the excess capital demand are both decreasing in r and w (at least locally around the equilibrium values).*

This means that the partial derivatives of $\Pi(r, w)$ and $X(r, w)$ with respect to r and w are negative:

$$\Pi_r < 0,\ \Pi_w < 0,\ X_r < 0 \text{ and } X_w < 0. \tag{A.14}$$

In a (w, r)-diagram, the slope of the iso-profit curve and of the iso-excess demand curve are given respectively by

$$S_\Pi = -\frac{\Pi_w}{\Pi_r} \text{ and } S_X = -\frac{X_w}{X_r}. \tag{A.15}$$

Therefore, equation (A.14) implies that both loci are downward sloping (i.e., $S_\Pi < 0$ and $S_X < 0$). In addition, a northeast shift of the locus (i.e., an increase in r and w) decreases the corresponding value of the respective iso-curve.

Furthermore, the relative position of the two loci is determined by the following property.

Property 3. *The gap between the two equilibrium conditions, $G \equiv X - \Pi$, is decreasing in r and increasing in w.*

Property 3 implies that the iso-profit curve is steeper than the iso-excess demand curve at all combination of (r, w) locally. To see this, note that the slopes of the iso-profit and iso-excess demand curves are given by equation (A.15). Since Property 3 indicates that the partial derivatives satisfy $G_r < 0$ and $G_w > 0$, we have

$$\Pi_r > X_r \ \text{ and } \ X_w > \Pi_w. \tag{A.16}$$

Therefore, the slopes of the two loci satisfy $|S_X| < |S_\Pi|$. A direct implication is the single-crossing property of the two loci: If the two curves ever cross they cross only once.[20] This establishes the uniqueness of the stationary equilibrium. Furthermore, the properties of the iso-curves indicate the direction for approaching the equilibrium from any off-equilibrium point.

[19] We thank Josef Falkinger for pointing out to us the basis of this section.

[20] The fact that the two curves cross (i.e., the existence of the equilibrium) is guaranteed in the numerical practice: In the region we search for the equilibrium, there always exist combinations of (r,w) on the zero-profit locus, s.t. $X(r, w) > 0$, and combinations, s.t. $X(r, w) < 0$. Since the zero-excess-capital-demand locus must lie between the loci that pass through the above mentioned two types of combinations, the zero-profit locus and the zero-excess-capital-demand locus cross.

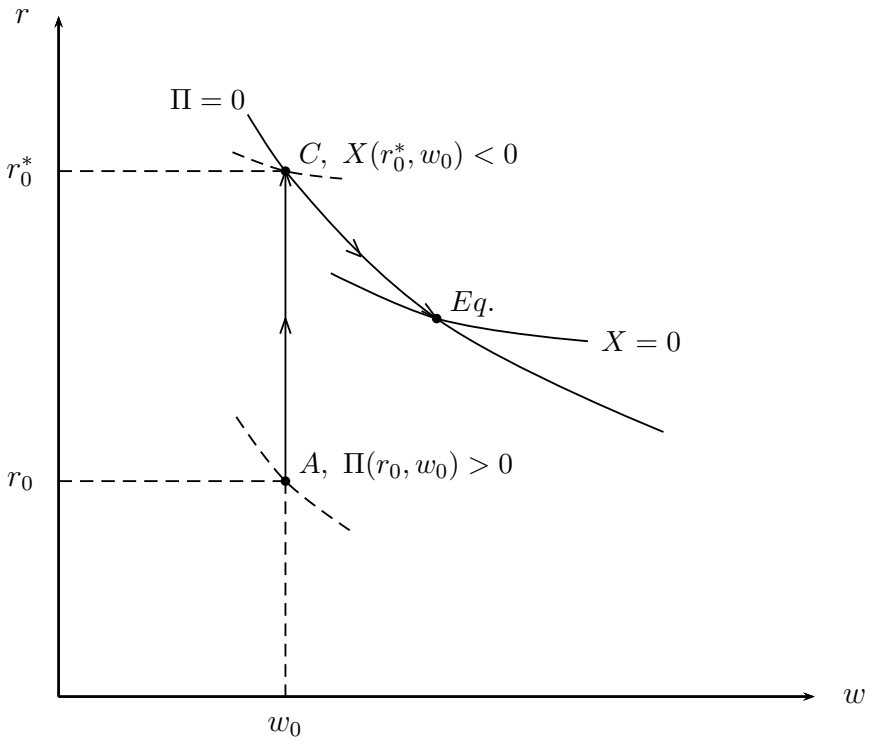

Figure A.2: Iso-profit and iso-excess demand curves

Notes: Note that we do not know the curvature of the two curves. Below the two solid lines profit and excess demand are positive and above they are negative.

Figure A.2 illustrates of iso-profit and iso-excess demand curves and gives an intuition of the algorithm to find the equilibrium, $Eq.$. Suppose that at an initial guess (r_0, w_0) (e.g., point A) the value of the iso-profit is $\Pi(r_0, w_0) > 0$. First, we approach the $\Pi = 0$ locus by changing r to r_0^*, s.t.(r_0^*, w_0) is on the locus (Step 4 and 5 in Algorithm A.3.4).[21] Then at (r_0^*, w_0) the excess capital demand $X(r_0^*, w_0)$ can be positive, negative or 0. In the last case we have found the equilibrium $Eq.$ directly. Now suppose $X(r_0^*, w_0) < 0$ (e.g., point C). Equation (A.14) and (A.16) suggest that the stationary equilibrium lies south-east of C.[22] Therefore, we shift (r, w) - $r \downarrow, w \uparrow$ - along the locus of the iso-profit curve $\Pi = 0$ until the excess demand increases to 0 ($C \rightarrow Eq.$).[23]

[21]Since zero is unachievable numerically, we use $|\Pi(r_0^*, w_0)| < \epsilon_P$ as a criterion for approximation. This applies for the the excess capital demand $X(r, w)$ as well. In addition, due to the unknown functional form of Π, it is impossible to calculate the exact increase in r_0 *ex ante* (i.e., $r_0^* - r_0$). This means that there may be back and forth in the adjustment of r. To guarantee that the target r_0^* is found in finite iterations, we record the upper and the lower bound of region where r_0^* lies in each iteration, and use binary search as is described in Step 5 and 8 in Algorithm A.3.4.

[22]This also means that $w_0 < w_{Eq}$. Therefore, w_0 is one lower bound of the equilibrium wage. We will update the lower bound if a new w', s.t. $w_0 < w' < w_{Eq}$ is found. The arguments apply for the upper bound as well.

[23]Similar to the situation described in footnote 21 it is not possible to find the correct adjustment in (r, w) in one step. Several iterations may be needed and we apply similar technique (i.e., recording upper and lower bounds and updating w in each iteration with binary search) to guarantee that the equilibrium

A.5 Intuition for convergence to stationary equity level

From the characteristics of the optimal contract (Figure 1.4), we notice that at low levels of promised values V^E expected repayments, $\pi_l m_l(V^E)+(1-\pi_l)m_h(V^E)$, from entrepreneurs to banks exceed the level of bank loans $b(V^E)$ and that the opposite holds at high levels of promised values (see Figure A.3). Intuitively, this means that banks receive a positive net cash flow from entrepreneurs with low promised values.

This positive net flow accrues to banks' equity. This is supplied as capital on the capital market and generates returns, which lead to a further accumulation of equity. In contrast, banks expect a negative net cash flow from firms with high promised values, which detracts banks' equity.

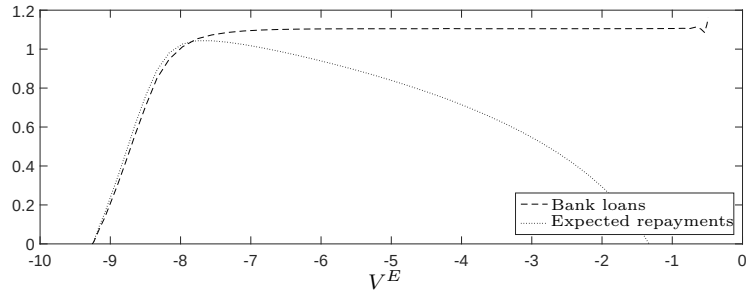

Figure A.3: Bank loans, $b(V^E)$, and expected repayments, $\pi_l m_l(V^E) + (1 - \pi_l)m_h(V^E)$

With this in mind, we can now intuitively describe the process of development of banks' equity level from the very beginning of time with no population to the stationary equity level E:[24] Suppose the banks are endowed with E_0 at the beginning of time when there is no population in the economy, yet. As population starts, there is a new-born cohort of entrepreneurs (and workers) with promised values $V^W(0; r, w) = V_0^E$. Entrepreneurs sign contracts with banks, which entitle them to banks loans and which ask for repayments. At the beginning of their lives, when entrepreneurs are at low levels of promised values they must give positive net cash flows to banks. Hence, banks start accumulating equity. With age, the average promised value of entrepreneurs increases (see firm dynamics in Figure 1.7 and 1.8) and reaches eventually levels where banks loans are larger than expected repayments. This reduces banks' equity. In addition, as the economy evolves, there are

is found in finite iterations. In addition, as we change (r, w) in each iteration, we need to make sure that the change is along the locus of the iso-profit curve $\Pi = 0$. Otherwise, we need to apply the first step again.

[24] Assume for simplicity that during the process of development interest rate and wage are fixed at some level (e.g., the equilibrium level (r, w)).

more overlapping cohorts – with younger cohort making positive and older cohorts making negative net cash flows to banks. In aggregation there is an accumulation of total bank's equity. Finally, in the stationary equilibrium the accumulation of banks' equity come to a halt so that the equity level stays constant. This means, in equilibrium negative aggregate net payments from entrepreneurs are exactly covered by the interest generated on banks' equity.

A.6 Figures

A.6.1 Illustration of productivity shock

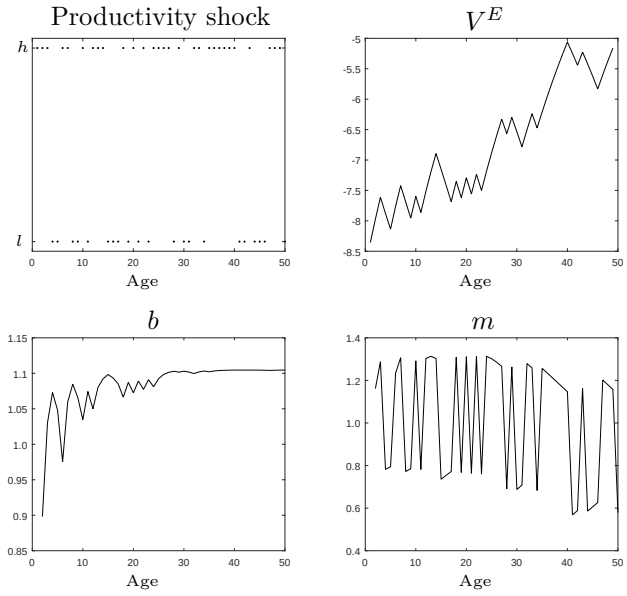

Figure A.4: Life path I

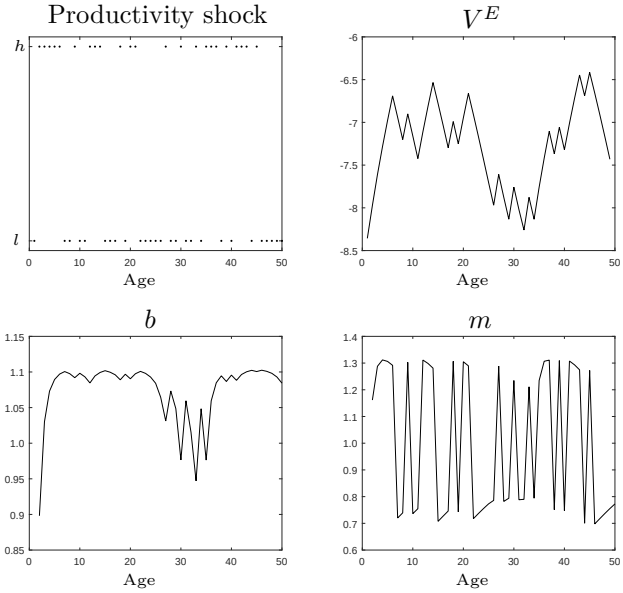

Figure A.5: Life path II

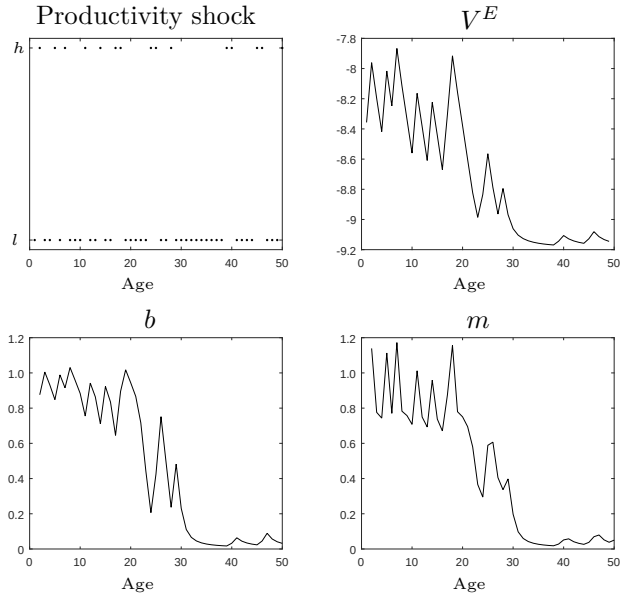

Figure A.6: Life path III

A.6.2 Development of entrepreneurs' bank loans and repayment

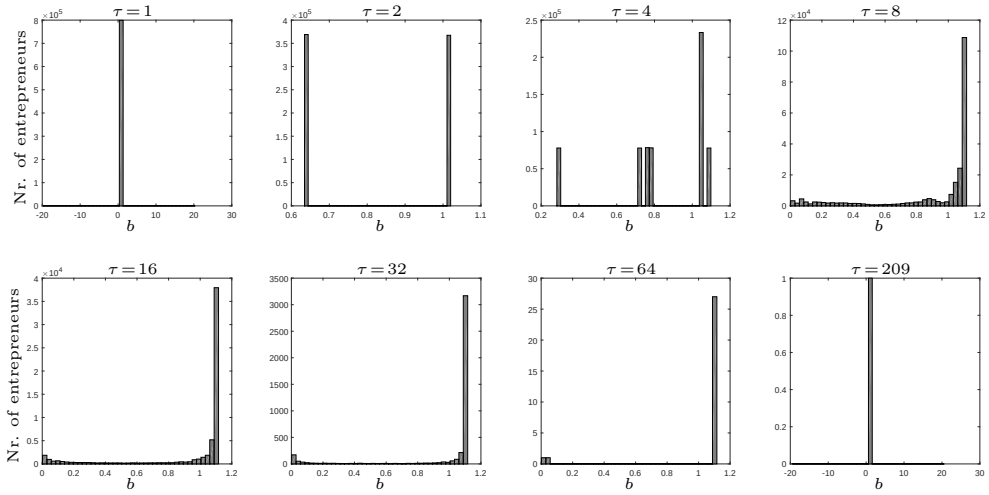

Figure A.7: Development of entrepreneurs' bank loans

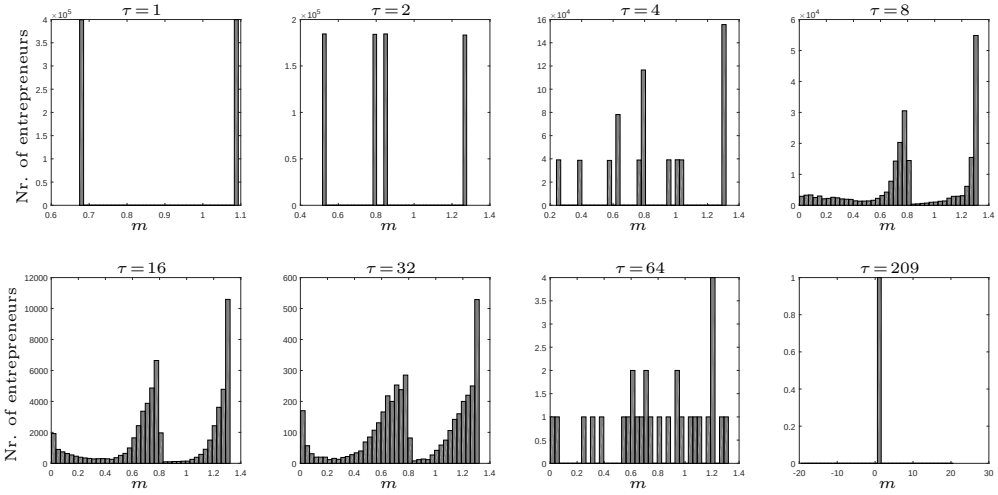

Figure A.8: Development of entrepreneurs' repayments to banks

Notes: Most of the subplots exhibit two distinct levels of repayments. This reflects the fact that within one cohorts firms may have high or low productivity realizations.

A.6.3 Production volatility

The quantitative results are summarized in Figure A.9.[25]

Subplot 1 shows the development of bank loans over lifetime under different levels of volatility, $\sigma \in \{0.15, 0.25, 0.375, 0.5\}$. As is discussed in Section 1.5.1.1, there exist simultaneously a volatility effect and an equilibrium price effect that changes the optimal

[25] We use again 5-year moving average after the same argument as in footnote 23.

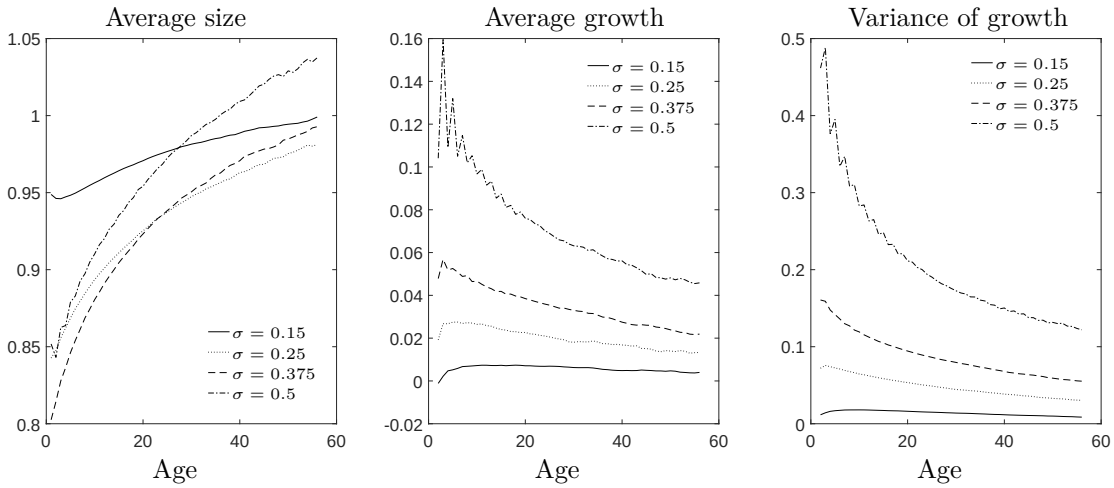

Figure A.9: Comparative statics of σ on firm dynamics

contract, and a distribution effect that changes the distribution of firms' promised values in the economy. Volatility effect decreases availability of credit in the economy, the firm sizes are smaller on average as production volatility increases (i.e., the curve tends to shift downwards). Similarly, distribution effect also has a negative impact on firms credit. However, the larger the volatility, the stronger the equilibrium price effect, which leads to an expansion of credit for all firms (i.e., the curve tends to shift upwards). Therefore, as volatility goes beyond the threshold (approximately $\bar{\sigma} = 0.0375$), the average bank loan increases (as can be seen by the fact that the curve with $\sigma = 0.5$ lies above the one with $\sigma = 0.375$). In addition, the volatility effect influences the credit availability of young firms more severely, which can be seen from the steeper slope as σ increases.

Subplots 2 and 3 show the average growth rate of firms and the variance of growth, respectively. At all ages the two values increase as production becomes more volatile. There are two effective channels: First, the efficient firm size defined in equation (1.35) is larger as σ increases. Since the optimal bank loan starts from $b = 0$ at $V^E = V_{min}$ and approaches the efficient level as V^E grows, the range of the bank loans is larger in more volatile economy. Second, the distribution of promised values is more dispersed as volatility increases, so is the distribution of bank loans. Therefore, the combination of two mutually amplifying channels leads to an increase of the growth rate and of the variance of growth.

A.6.4 Reserve ratio

Subplot 1 shows the development of bank loans over lifetime under different reserve ratios, μ. The equilibrium price effect and a lower initial promised values under high reserve ratios decrease the credit availability of young firms in the economy. However, average promised

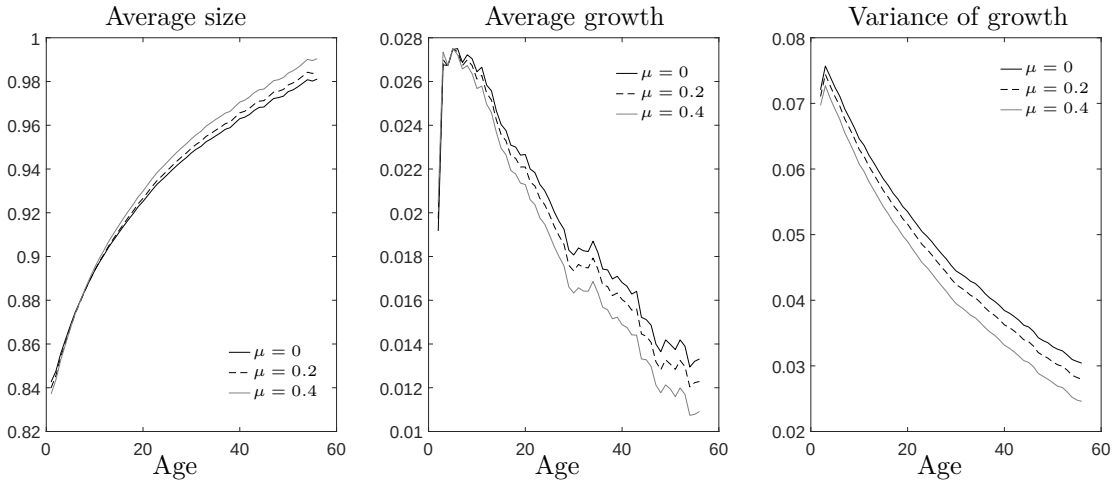

Figure A.10: Comparative statics of σ on firm dynamics

values increase as firms get older in an economy with higher reserve ratio. As a result, the size of old firms increases. The combination of the two gives the counter-clockwise shift of the average firm size as the reserve ratios increase.

Subplots 2 and 3 show that the average growth rate of firms and the variance of growth, respectively. As reserve ratios increase, both values decrease. This is mainly because the range of bank loans is smaller under a large reserve ratio: For a given change in the promised values, the change in the level of bank loans is lower. Therefore, the growth rate and variance of growth are smaller.

B Appendix: Chapter 2

B.1 Proofs

B.1.1 Portfolio Choice

Agent index l is skipped in the appendix. If financial intermediaries take ex-ante a fee in the form $T = p_{z_1} d + p_{z_2}(s-d)$, the expected utility maximization problem is given by:

$$\max_{s,\{f_\theta\}_{\theta\in\Theta},d} \mathbb{E}U = \log(e_0 - \bar{e}_0) + \delta \left[\mu \sum_{\theta\in\Theta} \pi_\theta \log(e_\theta - \bar{e}_1) + (1-\mu)\log(e_{\bar{\Theta}} - \bar{e}_1) \right]$$

s.t.

$$e_0 + (1+p_{z_2})s + (p_{z_1} - p_{z_2})d = y, \tag{B.1}$$

$$e_\theta = \begin{cases} R_\theta f_\theta + rd, \text{ if } \theta \in \Theta \\ rd, \text{ otherwise} \end{cases} \tag{B.2}$$

$$s = \sum_{\theta\in\Theta} f_\theta + d. \tag{B.3}$$

Denoting by λ the Lagrange multiplier for constraint (B.3) the first-order conditions of the households' expected utility maximization problem give:

$$\frac{\partial \mathcal{L}}{\partial s} = -\frac{1+p_{z_2}}{e_0 - \bar{e}_0} + \lambda = 0, \tag{B.4}$$

$$\frac{\partial \mathcal{L}}{\partial f_\theta} = \delta\mu\pi_\theta \frac{R_\theta}{e_\theta - \bar{e}_1} - \lambda = 0, \tag{B.5}$$

$$\frac{\partial \mathcal{L}}{\partial d} = -\frac{p_{z_1} - p_{z_2}}{e_0 - \bar{e}_0} + \delta \left[\mu \sum_{\theta\in\Theta} \pi_\theta \frac{r}{e_\theta - \bar{e}_1} + (1-\mu)\frac{r}{rd - \bar{e}_1} \right] - \lambda = 0, \tag{B.6}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = s - \sum_{\theta\in\Theta} f_\theta - d = 0. \tag{B.7}$$

Using (B.4), (B.5) and (B.6), we have

$$d = \frac{\delta(1-\mu)}{\lambda\left(\frac{1+p_{z_1}}{1+p_{z_2}} - r/R\right)} + \frac{\bar{e}_1}{r}. \tag{B.8}$$

where $R = \pi_\theta R_\theta$. From (B.2), (B.5) and (B.7), we have

$$s = \frac{\delta\mu}{\lambda} + (1 - r/R)d + \frac{1}{R}\bar{e}_1. \tag{B.9}$$

In the end we have

$$\begin{aligned} d &= \frac{\delta(1-\mu)}{(1+\delta)P}(y - \bar{e}_0) + \frac{(1+\mu\delta)(1+p_{z_1}) - (1+\delta)(1+p_{z_2})r/R}{r(1+\delta)P}\bar{e}_1 \\ &= \frac{1-\mu}{1-p\rho}\frac{\delta}{1+\delta}\frac{y-\bar{y}}{1+p_{z_1}} + \frac{\bar{e}_1}{r}, \end{aligned} \tag{B.10}$$

where $P \equiv (1+p_{z_1})(1-p\rho)$, $p \equiv \frac{1+p_{z_2}}{1+p_{z_1}}$, $\rho \equiv \frac{r}{R}$ and $\bar{y} \equiv \bar{e}_0 + \frac{\bar{e}_1(1+p_{z_1})}{r}$.

Combining (B.10) with (B.8) and solving for λ, we obtain

$$\frac{1}{\lambda} = \frac{y - \bar{y}}{(1+\delta)(1+p_{z_2})} \tag{*}$$

Using this and (B.10) in (B.9), we have

$$\begin{aligned} s &= \frac{\delta}{(1+\delta)}\frac{y-\bar{y}}{1+p_{z_2}}\left[\mu + (1-\rho)\frac{p(1-\mu)}{1-p\rho}\right] + (1-\rho)\frac{\bar{e}_1}{r} + \frac{\bar{e}_1}{R} \\ &= \frac{\delta}{1+\delta}\frac{y-\bar{y}}{1+p_{z_2}}\frac{\mu - p\rho + p(1-\mu)}{1-p\rho} + \frac{\bar{e}_1}{r}, \end{aligned}$$

which can be rewritten in the form

$$s = \frac{\delta}{1+\delta}\frac{y-\bar{y}}{1+p_{z_2}}\left[1 + \frac{(p_{z_2} - p_{z_1})(1-\mu)}{(1+p_{z_1})(1-p\rho)}\right] + \frac{\bar{e}_1}{r}, \tag{B.11}$$

where $p - 1 = \frac{p_{z_2} - p_{z_1}}{1+p_{z_1}}$ has been used.

Finally, (B.7), (B.10) and (B.11) give us

$$f \equiv \sum_{\theta\in\Theta} f_\theta = \frac{\mu - p\rho}{1-p\rho}\frac{\delta}{1+\delta}\frac{y-\bar{y}}{1+p_{z_2}} \tag{B.12}$$

and from (B.1) we conclude

$$\begin{aligned} y - e_0 &= (1+p_{z_1})d + (1+p_{z_2})f \\ &= \frac{\delta}{1+\delta}(y - \bar{y}) + \frac{(1+p_{z_1})\bar{e}_1}{r}. \end{aligned} \tag{B.13}$$

For the allocation of f on $f_\theta, \theta \in \Theta$, we combine (B.2) with (B.5) to get

$$\begin{aligned} f_\theta &= \pi_\theta \left[\frac{\delta\mu}{\lambda} + \frac{\bar{e_1} - rd}{R} \right] \\ &= \pi_\theta \frac{\delta}{1+\delta} \frac{y - \bar{y}}{1 + p_{z_2}} \left[\mu - \rho \frac{1-\mu}{1-p\rho} p \right] = \pi_\theta f, \end{aligned}$$

where (B.10) and (*) have been used for the second equation.

B.1.2 Corner solutions for securities demand

To account for the non-negativity constraint $f_\theta \geq 0$ we have to add $\sum_{\theta\in\Theta} \psi_\theta f_\theta$ to the Lagrange function for max EU – with $\psi_\theta \geq 0$ denoting the Lagrange multiplier for $f_\theta \geq 0$. Then, the first order condition for f_θ changes to

$$\delta\mu\pi_\theta \frac{R_\theta}{e_\theta - \bar{e}_1} - \lambda + \psi_\theta = 0 \tag{B.14}$$

with $\psi_\theta f_\theta \leq 0$.

Suppose that $f_\theta = 0$ for all θ. Then $s = d$ and

$$\begin{aligned} e_0 - \bar{e}_0 &= y - \bar{e}_0 - (1 + p_{z_1})d \\ e_\theta - \bar{e}_1 &= rd - \bar{e}_1 \end{aligned} \tag{B.15}$$

and the first-order conditions

$$\begin{aligned} (s) \qquad & \lambda = \frac{1 + p_{z_2}}{e_0 - \bar{e}_0} \\ (d) \qquad & \delta \left[\mu \sum_{\theta\in S} \pi_\theta \frac{r}{e_\theta - \bar{e}_1} + (1-\mu) \frac{r}{rd - \bar{e}_1} \right] = \lambda + \frac{p_{z_1} - p_{z_2}}{e_0 - \bar{e}_0} \end{aligned} \tag{B.16}$$

reduce to

$$\delta \frac{r}{rd - \bar{e}_1} = \frac{1 + p_{z_1}}{e_0 - \bar{e}_0}.$$

With (B.15) this solves to

$$d = \frac{1}{1+\delta} \left[\frac{\delta(y - \bar{e}_0)}{1 + p_{z_1}} + \frac{\bar{e}_1}{r} \right]. \tag{B.17}$$

Substituting the solution into (B.15) gives us

$$\begin{aligned} e_0 - \bar{e}_0 &= \frac{1}{1+\delta}\left[y - \bar{e}_0 - \frac{(1+p_{z_1})\bar{e}_1}{r}\right] \\ e_\theta - \bar{e}_1 &= \frac{\delta r}{(1+\delta)}\left[\frac{y-\bar{e}_0}{1+p_{z_1}} - \frac{\bar{e}_1}{r}\right]. \end{aligned} \tag{B.18}$$

Using this in (B.14) we obtain: $\psi_\theta \geq 0$ if and only if

$$\mu\pi_\theta R_\theta \leq \frac{1+p_{z_2}}{1+p_{z_1}} r \tag{B.19}$$

where $\lambda = \frac{1+p_{z_2}}{e_0-\bar{e}_0}$ has been used from (B.16).

Since $\pi_\theta R_\theta = R$, (B.19) reduces to

$$\frac{1+p_{z_1}}{1+p_{z_2}}\mu R \leq r,$$

which is equivalent to $R\mu(1+p_{z_1}) \leq (1+p_{z_2})r$.

Hence non-negativity $f_\theta > 0$, $\theta \in \Theta$, requires

$$R\mu(1+p_{z_1}) > (1+p_{z_2})r. \tag{B.20}$$

B.1.3 Further proofs

Proof of Fact 2.2. With (2.11) and (2.12) the condition $y^L = b_L w_L > \bar{y} = \bar{e}_0 + \frac{(1+p_z)\bar{e}_1}{r}$ takes the form

$$A_x\Gamma_x\omega^{-\alpha_x}\left[b_L - \frac{\bar{e}_1}{rA_{z_1}\Gamma_{z_1}}\omega^{\alpha_{z_1}}\right] > \bar{e}_0 + \frac{\bar{e}_1}{r}.$$

The left side of the equation declines in ω. Thus $y^L > \bar{y}$ requires

$$\omega < \omega_L^+\left(\underset{+}{A_x}, \underset{+}{A_{z_1}}, \underset{+}{b_L}, \underset{-}{\bar{e}_0}, \underset{-}{\frac{\bar{e}_1}{r}}\right),$$

where ω_L^+ is determined by the equation:

$$b_L = (\bar{e}_0 + \frac{\bar{e}_1}{r})\frac{\omega^{\alpha_x}}{A_x\Gamma_x} + \frac{\bar{e}_1}{r}\frac{\omega^{\alpha_{z_1}}}{A_{z_1}\Gamma_{z_1}}.$$

□

Proof of Lemma 2.1. a) Let $B_1 \equiv A_x\Gamma_x\frac{b_L\bar{L}}{N}$ and $B_2 \equiv \frac{A_x\Gamma_x}{A_z\Gamma_z}$. Using (2.21) and (2.12), we

have

$$\bar{w} = B_1\omega^{-\alpha_x}(1+\omega k), \quad p_z = B_2\omega^{\alpha_z-\alpha_x}.$$

Then $\bar{\eta}$ can be reformulated as

$$\bar{\eta} = \frac{\bar{w}-\bar{y}}{1+p_z} = \frac{B_1\omega^{-\alpha_x}(1+\omega k)-\bar{e_0}}{1+B_2\omega^{\alpha_z-\alpha_x}} - \frac{\bar{e_1}}{r},$$

where (2.14) is used to substitute $\bar{y}$.

To get the shape of $\bar{\eta}$, first notice that

$$sign\frac{\partial\bar{\eta}(\omega)}{\partial\omega} = sign\frac{\partial G(\omega)}{\partial\omega},$$

where $G(\omega) \equiv \frac{B_1(1+\omega k)-\bar{e_0}\omega^{\alpha_x}}{\omega^{\alpha_x}+B_2\omega^{\alpha_z}}$. Differentiating $G(\omega)$ we have

$$\frac{\partial G(\omega)}{\partial\omega} = \frac{\mathcal{L}(\omega)}{(\omega^{\alpha_x}+B_2\omega^{\alpha_z})^2},$$

where

$$\begin{aligned}\mathcal{L}(\omega) =& B_1\omega^{\alpha_x}\left[k(1-\alpha_x)-\frac{\alpha_x}{\omega}\right] + B_1B_2\omega^{\alpha_z}\left[k(1-\alpha_z)-\frac{\alpha_z}{\omega}\right] \\ &+ \bar{e_0}B_2(\alpha_z-\alpha_x)\omega^{\alpha_x+\alpha_z-1}.\end{aligned}$$

We have $\frac{\partial G(\omega)}{\partial\omega} > 0$ if and only if $\mathcal{L}(\omega) > 0$. For $\alpha_x+\alpha_z > 1$, $\mathcal{L}(\omega)$ is an increasing function in ω. Moreover,

$$\lim_{\omega\to 0^+}\mathcal{L} = -\infty, \quad \lim_{\omega\to+\infty}\mathcal{L} = +\infty.$$

Therefore, there exists a unique $\underline{\omega}$ with $\mathcal{L}(\underline{\omega}) = 0$ and: $\frac{\partial\bar{\eta}(\omega)}{\partial\omega} \gtreqless 0$ if and only if $\omega \gtreqless \underline{\omega}$. A rise in k or $\bar{e}_0$ shifts $\mathcal{L}(\omega)$ upward so that $\underline{\omega}$ declines. The impacts of B_1, B_2 (and thus of A_x, A_z, $\frac{b_L\bar{L}}{N}$) on $\underline{\omega}$ are ambiguous because $\kappa_x < k < \kappa_z$ imply $k(1-\alpha_x)-\frac{\alpha_x}{\omega} > 0$ and $k(1-\alpha_z)-\frac{\alpha_z}{\omega} < 0$.

b) We have

$$\bar{\eta} = \frac{A_x\Gamma_x\frac{b_L\bar{L}}{N}\omega^{-\alpha_x}(1+\omega k)-\bar{e_0}}{1+\frac{A_x\Gamma_x}{A_z\Gamma_z}\omega^{\alpha_z-\alpha_x}} - \frac{\bar{e_1}}{r}.$$

By eye inspection we get:

$$\bar{\eta}\left(\omega\middle|\underset{+}{A_x},\underset{+}{A_z},\underset{+}{k},\underset{+}{\frac{b_L\bar{L}}{N}},\underset{-}{\bar{e_0}},\underset{-}{\frac{\bar{e_1}}{r}}\right)$$

□

Proof of Fact 2.5. According to (2.32), $Z^S = A_z b_L \bar{L} \frac{\gamma_z^{\alpha_z}}{\gamma_z - \gamma_x} \omega^{-\alpha_z}(k\omega - \gamma_x)$, where $\kappa_j = \frac{\gamma_j}{\omega}$ has been used from (2.9).

We have $\frac{\partial \omega^{-\alpha_z}(k\omega - \gamma_x)}{\partial \omega} = \omega^{-\alpha_z}\left[(1-\alpha_z)k + \frac{\alpha_z \gamma_x}{\omega}\right]$. This term is positive and decreasing in ω. □

B.2 Extensions

Five extensions are considered: Fixed costs in the financial sector, rents in the financial sector, distorted portfolio choices of households, participation constraints in finance subsector Z_2 and set-up capital for firms. Like the equilibrium analysis in the benchmark, the extended analysis is based on Assumption 2.2. Moreover, for avoiding too many case distinctions, dominance of $\bar{e}_0$ over $\bar{e}_1$ is assumed in this section.

B.2.1 Fixed costs in the financial sector

Suppose that financial services are provided by banks. A bank b, serving N_b clients, needs $K_b = f_B N_b$ units of goods to set up the capacity to serve them. We assume that the fixed cost K_b is financed by a lump-sum fee

$$\tau = f_B$$

imposed on the clients. That is, bank size and number of banks affect neither aggregate fixed costs

$$K_B = f_B N$$

nor the households' budget constraint. In the latter, y^l reduces to $y^l - \tau$ so that the supernumerary budget becomes $y^l - \bar{y}_+$, with $\bar{y}_+ = \bar{y} + \tau = \bar{e_0} + f_B + (1 + p_{z_1})\bar{e_1}/r$.

Hence, fixed cost f_B has the same comparative-static effects on household choices as an increase in subsistence expenditure $\bar{e_0}$. For the X-market this means, on the one hand, the absorption of X by households' consumption and investment is reduced by $K_B = f_B N$. On the other hand, K_B is spent by banks to set up the capacity to serve their clients. In sum, we have

$$E_0 - f_B N + D + F + K_B = X$$

for the goods market clearing, which reduces to the condition in the benchmark model:

$$E_0 + D + F = X$$

since $f_B N = K_B$. Hence, goods markets are cleared whenever the Z-markets are cleared.

In the markets for financial services, demand is reduced by the fact that $\bar{w} - \bar{y}_+$ rather than $\bar{w} - \bar{y}$ is now the relevant supernumerary income. The supply side remains unaffected. In equilibrium, the implications of fixed costs can be derived by looking in the benchmark model at the effect of a rise of $\bar{e_0}$ to $\bar{e_0} + f_B$.

Proposition B.1. *A decline in fixed costs f_B has the following effects:*

a) The skill premium rises.

b) The between sectoral structure shifts from X to Z.

c) The within sectoral structure shifts from Z_1 to Z_2 at high levels of the skill premium ($\omega^* > \underline{\omega}$). At low levels of the skill premium ($\omega^* < \underline{\omega}$) the effect is ambiguous.

Proof. Comparative-static results for $\bar{e_0}$ in Proposition 2.4, 2.5 and 2.6. □

B.2.2 Rents in the financial sector

Suppose that a club of agents in the finance sector has the power to extract rents from financial service provision.[1] One may think of rentiers who have unearned property rights or an elite subgroup of employees in the financial sector. We make two crucial assumptions: First, whoever are the rent extracting agents, they spend the rent like other agents. Thus, the redistribution of rents has no income effect on aggregate demand. (Total subsistence requirements and aggregate supernumerary income remain unchanged). Second, nobody can enter the club from outside so that the rent does not affect labor allocation.

In the presented model, two instruments can be used to extract rents. First, a fixed fee $\tilde{\tau}$ as in extension B.2.1, but:

$$\tilde{\tau} > f_B.$$

Aggregate rents $(\tilde{\tau} - f_B)N$ are lump-sum redistributed. Everybody pays $\tilde{\tau}$ and an elite N_0 receives the rent. Thus, average supernumerary income becomes

$$\bar{w} - \bar{y} - \tilde{\tau} + \frac{N_0}{N}\frac{(\tilde{\tau} - f_B)N}{N_0} = \bar{w} - \bar{y} - f_B.$$

In this case, the rent has no effects on aggregate income, expenditure structure, labor allocation, relative prices or the skill premium. Nevertheless, there is lump-sum redistribution of income from the real to the financial sector and within the financial sector. This

[1] As pointed out in the introduction, there is robust evidence that indeed a substantial finance premium exists. This paper deals with the consequences of rents, not with possible explanations why they exist.

redistribution implies for the sectoral income shares:

$$\frac{p_z Z + (\tilde{\tau} - f_B)\, N}{X}$$

and

$$\frac{p_z F + \nu(\tilde{\tau} - f_B)N}{p_z D + (1-\nu)(\tilde{\tau} - f_B)N},$$

respectively, where ν is the share of the elite rent going to new finance. It is obvious that a rising finance rent increases the total finance share in the economy. For a given rent distribution ν, a rise in $\tilde{\tau}$ raises the income share of new finance relative to traditional finance as long as $\nu D > (1-\nu)F$, that is as long as the new finance share is not too large. A rise in ν trivially leads to a rise in the new finance share.

A second instrument of rent extraction would be to charge a markup on unit cost prices in the financial sector so that households have to pay $\tilde{p}_{z_i} = p_{z_i}(1 + o_i)$ for financial services.

Using (2.12), we have

$$\tilde{p}_{z_i} = (1 + o_i)\frac{A_x}{A_{z_i}}\frac{\Gamma_x}{\Gamma_{z_i}}\omega^{\alpha_{z_i} - \alpha_x}.$$

In the benchmark case with $p_{z_1} = p_{z_2}$ a rent $o_1 = o_2 = o$ decreases D_1 in (2.36) to $\frac{A_x \Gamma_x (1+\omega k)}{\omega^{\alpha_x} + \frac{(1+o) A_x \Gamma_x}{A_z \Gamma_z}\omega^{\alpha_z}}$ and decreases D_0 in (2.37) to $\frac{1}{1+\delta}\left[\frac{\delta \bar{e_0}}{1+(1+o)\frac{A_x \Gamma_x}{A_z \Gamma_z}\omega^{\alpha_z - \alpha_x}} - \frac{\bar{e_1}}{r}\right]$. Hence, o has an ambiguous impact on $Z_D - Z_S$ and thus on ω^*.

Proposition B.2. *Rents in the financial sector have the following effects:*

a) If rents are extracted by lump sum fees, they have no allocative equilibrium effects. Yet, there is a redistributive effect that raises the finance share in total income. The structure of the subsector shares within finance depends on how the earned rents are distributed on traditional and new finance, respectively.

b) If rents are extracted by a markup on financial service prices, there is a redistributive effect towards (and within) the financial sector. Yet, the mark ups affect all equilibrium values in a generally ambiguous way.

Proof. Main text. □

B.2.3 Distorted portfolio choice

Several empirical studies have pointed out that people get confused in dealing with complex financial markets (see Célérier and Vallée (2014) and the literature discussed there). In our model, the complex part that households have to solve is the choice of the portfolio of the securities. The choice may be based on a wrong assessment of relative risks and

returns of different securities. In this case, we have distortion within Z_2 and consumption levels planned for the future may be deceived by actual payoffs.[2] As our study focuses on structural change between X and Z as well as between Z_1 and Z_2, we do not consider such distortions here. Rather we focus on distortions coming from misperception of the opportunities to save by securities investment rather than in deposits.

In particular, people may have wrong beliefs $\tilde{\mu}$ about the measure of future environments covered by state-contingent securities, relative to the non-covered part of possible future events. They may also misjudge the relative payoff of deposits compared to the payoffs of securities and base their decisions on a distorted $\tilde{\rho}$. Such distortions affect the propensities to save in deposits and in securities. For instance, if agents are euphoric about investments in securities and believe that $\tilde{\mu} > \mu$ or $\tilde{\rho} < \rho$, then s_f rises while s_d declines. The total propensity to save, however, does not change in the benchmark model with $p_{z_1} = p_{z_2}$.[3] Therefore, the only consequence of $\tilde{\mu} > \mu$ or $\tilde{\rho} < \rho$ is sectoral change within the financial sector. According to (2.27), Φ rises.

Proposition B.3. *Euphoric beliefs about measure or performance of state-contingent financial instruments lead to within sectoral change from Z_1 to Z_2. Equilibrium skill premium and (X, Z)-structure are not affected in the benchmark model (with identical technologies in Z_1 and Z_2).*

Proof. Equation (2.27). □

B.2.4 Participation constraints

Suppose that a fixed fee τ is charged only to agents who invest in securities. Moreover, assume that there is a participation constraint:

$$y^L > \bar{y} > y^L - \tau,$$
$$y^H > y^H - \tau > \bar{y}.$$

Then low-skilled agents do not participate in the securities market, while high-skilled agents do. According to equation (B.17) in Appendix B.1.2, we have for $l = L$:

$$s^L = d^L = \frac{\delta}{1+\delta}\frac{y^L - \bar{y}}{1+p_z} + \frac{\bar{e}_1}{r}.$$

For $l = H$, saving behavior is given by (2.16) and (2.17) with $\bar{y}_+ = \bar{y} + \tau$.

[2] Falkinger (2014) focuses on such distortions in a one sector economy.

[3] For $p_{z_1} \neq p_{z_2}$, however, we would have $s_d + \frac{s_f}{p}$ for the marginal propensity to save. Thus, μ and ρ impact also on Z^D and therefore on ω and all other equilibrium outcomes. See Section B.3 for a more detailed discussion.

This gives us the following aggregate saving levels:

$$\begin{aligned} D &= \frac{\delta}{1+\delta}\frac{1}{1+p_z}\left[(y^L-\bar{y})\bar{L}+s_d(y^H-\bar{y}_+)\bar{H}\right]+\frac{\bar{e_1}}{r}N \\ F &= s_f\frac{\delta}{1+\delta}\frac{\bar{H}}{1+p_z}(y^H-\bar{y}_+) \\ S &= \left(\frac{\delta}{1+\delta}\frac{\bar{w}-\bar{y}-\tau\frac{\bar{H}}{N}}{1+p_z}+\frac{\bar{e_1}}{r}\right)N. \end{aligned}$$

Comparing S with Z^D in (2.31), we see that fee τ, combined with the participation constraint, impacts on Z^D and thus on the skill premium and the (X,Z)-structure like an increase of $\bar{e_0}$ to

$$\tilde{e_0} = \bar{e_0}+\tau\frac{\bar{H}}{N}.$$

Moreover, $\frac{F}{D} = \frac{s_f\bar{H}}{\frac{(y^L-\bar{y})\bar{L}}{y^H-\bar{y}_+}+s_d\bar{H}+\frac{1+\delta}{\delta}\frac{(1+p_z)\bar{e}_1}{r}\frac{N}{y^H-\bar{y}_+}}$ is declining in τ. Thus, the participation constraint does not change the comparative static effects of fixed cost τ described in Proposition B.1.

The above conclusion is only valid if τF is absorbed by real fixed cost requirements as discussed in Section B.2.1. If τF is a rent which is redistributed back to high-skilled agents, we have $(y^H-\bar{y}-\tau)\bar{H}+\tau\bar{H} = y^H-\bar{y}$ instead of $y^H-\bar{y}_+$ so that

$$\begin{aligned} D &= \frac{\delta}{1+\delta}\frac{\bar{w}-\bar{y}}{1+p_z}N\left(1-s_f\beta_H\right)+\frac{\bar{e_1}}{r}N \\ F &= s_f\frac{\delta}{1+\delta}\frac{\bar{H}}{1+p_z}(y^H-\bar{y}) \\ S &= \left(\frac{\delta}{1+\delta}\frac{\bar{w}-\bar{y}}{1+p_z}+\frac{\bar{e_1}}{r}\right)N \end{aligned}$$

with $\beta_H \equiv \frac{y^H-\bar{y}}{\bar{w}-\bar{y}}\frac{\bar{H}}{N}$ denoting the income share of high-skilled agents. For the high-skilled nothing changes, but the low-skilled are only saving through D. This means that, compared to the benchmark, we have an increase in D and a decrease in F. $Z^D = S$ coincides with the expression in (2.31) so that equilibrium skill premium and (X,Z)-structure are not changed compared to the baseline.[4]

For the within sectoral structure in the Z-sector, we have in the benchmark case with

[4]For $p_{z_1} \neq p_{z_2}$, however, the change in Z_2^D would also affect ω and all other equilibrium outcomes.

$p_{z_1} = p_{z_2} = p_z$:

$$\frac{F}{D} = \frac{s_f \beta_H}{1 - s_f \beta_H + \frac{1+\delta}{\delta} \frac{1+p_z}{\bar{w}-\bar{y}} \frac{\bar{e_1}}{r}}$$
$$= \frac{s_f \beta_H \bar{\eta}}{s_d \bar{\eta} + s_f (1 - \beta_H) \bar{\eta} + \frac{1+\delta}{\delta} \frac{\bar{e_1}}{r}} \equiv \tilde{\Phi}$$

Comparing this with (2.27), we conclude that $\tilde{\Phi} < \Phi$ because $s_f(1 - \beta_H) > 0$. Yet, the proportion of total expenditure on new finance relative to expenditure on traditional finance $\frac{p_z F + \tau \bar{H}}{p_z D} = \frac{F}{D} + \frac{\tau \bar{H}}{p_z D}$ is ambiguous. Rent τ increases the new finance share, but the participation constraint induces a shift of the portfolio towards safe assets.

B.2.5 Set-up capital for firms

In the baseline model invested capital is transformed by linear technologies, using capital as the only input, into future outcome. The extension in this section shows that the baseline can be seen as kind of reduced form of a richer model, in which capital is needed to set up firms. We assume now that firms in the X-sector use capital to set up technology G^x, which then produces output by employing low-skilled and high-skilled labor. Each established firm $\nu \in \{1, ..., M\}$ produces a variety $x_\nu = G^x(L_{x_\nu}, H_{x_\nu})$ under monopolistic competition with free entry. Consumers spend the supernumerary income $e_t - \bar{e_t}$ according to a CES-utility function with substitution elasticity $\sigma > 1$ symmetrically over the variants x_ν in the X-sector, which implies an instantaneous indirect utility function of the form $\log(e_t - \bar{e_t})$ (see Section B.2.5.1) like before. So saving decision and portfolio choice remain the same as in the baseline model. Firms have positive operating profits which are distributed as payoff to the investors (see Section B.2.5.2).

B.2.5.1 Consumer problem

Let the instantaneous utility of households be given by $u = \left[\sum_{\nu=1}^{M} x_\nu^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}}$, $\sigma > 1$. Then, prices are determined by a constant markup on unit cost of production

$$p_\nu = \frac{\sigma}{\sigma - 1} c(w_H, w_L), \tag{B.21}$$

where $c(w_H, w_L)$ are the unit costs (as in Section 2.3) and w_H, w_L are factor prices. Moreover, demand for variety x_ν of a household that spends "supernumerary budget"

$e - \bar{e}$ is

$$x_\nu = (e - \bar{e})\frac{p_\nu^{-\sigma}}{P^{1-\sigma}}, \quad P \equiv \left[\sum_{\nu=1}^{M} p_\nu^{1-\sigma}\right]^{\frac{1}{1-\sigma}}.$$

Since product variants use identical production technologies, their unit cost and prices are identical, too. Thus, x_ν reduces to $x = \frac{e-\bar{e}}{p_\nu M}$. Using this in u, we obtain for the instantaneous indirect utility $u = \frac{e-\bar{e}}{P}$. We set the price as numéraire (i.e., $p_\nu = 1$) so that the variety effect is $P = M^{\frac{1}{1-\sigma}}$. Due to the log specification, this variety effect, though affecting the level of utility, does not matter for the intertemporal decision.[5] Thus, $\max \mathbb{E}log(u) = \max \mathbb{E}log(e_t - \bar{e}_t)$, which is identical to the intertemporal problem in Section 2.3.2.

B.2.5.2 Firm entry and production in the X-sector

There are two types of set-up technologies, which induce capital demand of firms: A robust set-up technology which requires c_0 units of capital. Firms set up by the robust technology will be producing tomorrow under any condition (i.e., in Θ and $\bar{\Theta}$). Furthermore, there are risky set-up technologies with set-up input c_θ, which are only effective if state $\theta \in \Theta$ occurs. Otherwise, their set-up fails. In an analogous way to (2.1), we assume

$$c_\theta = \pi_\theta c_1, \text{ where } c_1 < c_0. \tag{B.22}$$

The assumption states that set-up capital required for a robust technology is larger than the capital required for risky technologies. Moreover, the smaller the measure π_θ of the state under which a set-up technology works, the lower the required set-up capital.[6] Robust set-up technologies are financed by loans, whereas the risky set-up techniques are financed by state-contingent securities.

Let K_0 be the aggregate set-up capital for robust technologies and denote by $K_\theta, \theta \in \Theta$, the aggregate set-up capital for specialized risky technologies. Then the number of firms which can be set up is $M_0 = \frac{K_0}{c_0}$ and $M_\theta = \frac{K_\theta}{c_\theta}$, respectively. In a closed economy, capital markets are cleared if

$$K_0 = D, \quad K_\theta = F_\theta = \pi_\theta F.$$

[5] Note that $log\frac{e-\bar{e}}{P} = log(e - \bar{e}) - logP$ so that the P-levels add to $\mathbb{E}U$ a constant.

[6] See Falkinger (2014) for a more detailed discussion of the relationship between specialization and risk. There, technologies are more productive the more narrowly they are targeted to a specific environment. At the same time, they are more risky because the realization of the specific environment is less likely. Here this idea is applied to set-up costs rather than productivity.

Hence, we have for to total number of firms

$$M = \begin{cases} \frac{D}{c_0} + \frac{F}{c_1} \equiv M_\Theta, & \text{if } \theta \in \Theta, \\ \frac{D}{c_0} \equiv M_{\bar{\Theta}} & \text{otherwise.} \end{cases}$$

After firms being set up, their operating profits earned under mark-up prices (B.21) are

$$\Pi = (p_x - c)X = \frac{X}{\sigma},$$

where $p_x = 1$, which implies $c = \frac{\sigma-1}{\sigma}$, has been used. Since firms are symmetric, aggregated operating profits are distributed uniformly across firms so that operating profit per firm is:

$$\frac{\Pi_m}{M_m} = \frac{X}{\sigma M_m}, \ m \in \{\Theta, \bar{\Theta}\}.$$

The returns on one unit of set up capital are therefore

$$r_m = \frac{X}{c_0 \sigma M_m}, \ m \in \{\Theta, \bar{\Theta}\}$$
$$R_\theta = \frac{X}{c_\theta \sigma M_\Theta}, \ R = \frac{X}{c_1 \sigma M_\Theta}$$

for safe and risky investments, respectively. ($\pi_\theta R_\theta$ reduces to R because of assumption $c_\theta = \pi_\theta c_1$.) Since the number of firms is different in Θ and $\bar{\Theta}$, aggregate operating profits have to be shared among more or less firms so that the return on robust investments is m-dependent. The relative rate of return, $\frac{r_\Theta}{R_\theta}$, however, is uniquely determined by the relative set-up requirements of specialized risky technologies compared to the robust technology. We have $\rho = \frac{c_1}{c_0}$.

For the portfolio choice derived in Section 2.5 almost only the relative rate ρ matters. The exception is $\frac{\bar{e}_1}{r_m}$, since future subsistence can only be financed by deposits.[7] This means, we have to restrict the analysis of the paper to $\bar{e}_1 = 0$, or we reconcile the fluctuation of the earnings of robust firms with a safe return on deposits by assuming that firms hold buffers and distribute the expected profit per firm $\bar{\pi} \equiv [\frac{\mu}{M_\Theta} + \frac{1-\mu}{M_{\bar{\Theta}}}]\frac{X}{\sigma}$ to the investors.

For the general equilibrium analysis, a further caveat is in order. Under the presented extension, return r (even if smoothed by the buffer) is endogenous. It depends on M

[7] Formally the derivation of the portfolio choice presented in the appendix has to be adapted to account for m-dependent pay-offs in the budget constraints. For $\bar{e}_1 = 0$, return $r_{\bar{\Theta}}$ becomes irrelevant under the logarithm specification and the analysis remains valid – with $\rho = \frac{r_\Theta}{R_\Theta}$.

and X, which are determined by saving behavior and resource allocation, respectively. Thus, in the general equilibrium, a further feedback loop is to be considered. We did not account for such feedbacks in Section 2.7, since in the baseline return r is exogenously given by the constant productivity of capital. For $\bar{e}_1 = 0$, however, the presented analysis remains fully valid also with set-up capital of firms, since r matters only through the term $\frac{\bar{e}_1}{r}$. However, what one loses by setting $\bar{e}_1 = 0$ is the income effect on structural change within the financial sector. For the income effect on the skill premium and the structural change between goods and financial sector subsistence level $\bar{e}_0 > 0$ is relevant, which poses no problem in the extension considered here.

B.3 Robustness

To account for relative price effects within the financial sector, we skip now Assumption 2.2 and impose the following restriction instead.

Assumption 1.2'. $\alpha_x = \alpha_{z_1} < \alpha_{z_2}$.

Then, according to (2.12),

$$p_{z_1} = \frac{A_x}{A_{z_1}}$$

and thus: $\bar{y} = \bar{e_0} + \frac{(1+\frac{A_x}{A_{z_1}})\bar{e_1}}{r}$.

Moreover, the terms $a_x^l X + a_{z_1}^l Z_1$, $l \in \{H, L\}$, in system (2.13) reduce to

$$X^+ \frac{1}{A_x \kappa_x^{\alpha_x}} \text{ and } X^+ \frac{\kappa_x^{(1-\alpha_x)}}{A_x},\ X^+ \equiv X + \frac{A_x}{A_{z_1}} Z_1,$$

respectively. Using this when solving (2.13), we obtain

$$X^+ = \frac{b_L \bar{L}}{a_x^L} \frac{\kappa_{z_2} - k}{\kappa_{z_2} - \kappa_x},\ Z_2 = \frac{b_L \bar{L}}{a_{z_2}^L} \frac{k - \kappa_x}{\kappa_{z_2} - \kappa_x} \tag{B.23}$$

and

$$\frac{p_{z_2} Z_2}{X^+} = \frac{p_{z_2}(\omega) a_x^L(\omega)}{a_{z_2}^L(\omega)} \frac{k - \kappa_x(\omega)}{\kappa_{z_2}(\omega) - k} \equiv \tilde{\Psi}(\underset{+}{\omega}, \underset{+}{k}) \tag{B.24}$$

where the signs for the partial derivatives of $\tilde{\Psi}$ follow from $\kappa_{z_2} > k > \kappa_x$, the Rybczynski analysis and the fact that p_{z_2} rises in ω.

Substituting $A_{z_2} \kappa_{z_2}^{\alpha_{z_2}}$ for $\frac{1}{a_{z_2}^L}$ in the second equation of (B.23) and using (2.9), we have for the Z_2-supply:

$$Z_2^S = A_{z_2} b_L \bar{L} \frac{\gamma_{z_2}^{\alpha_{z_2}}}{\gamma_{z_2} - \gamma_x} g(\underset{+}{\omega}, \underset{+}{k}), \quad g(\omega, k) \equiv \omega^{-\alpha_{z_2}}(k\omega - \gamma_x). \tag{B.25}$$

This coincides with (2.34) – with Z_2 instead of Z – so that Fact 2.5 remains valid under the alternative specification and applies to Z_2-supply.

Z_2-demand is given by

$$Z_2^D = F = s_f \frac{\delta}{1+\delta} \frac{\bar{w}-\bar{y}}{1+p_{z_2}} N = \frac{\mu-\rho p}{1-\rho p} \frac{\delta}{1+\delta} \tilde{\eta} N \tag{B.26}$$

with $\tilde{\eta} \equiv \frac{\bar{w}-\bar{y}}{1+p_{z_2}}$ and $p = \frac{1+p_{z_2}}{1+p_{z_1}}$. In an analogous way to Lemma 2.1 and Fact 2.6, one establishes that the income effect (i.e., $\tilde{\eta}$-part in Z_2^D) has an U-shaped form.[8] Further, s_f is decreasing in ω since $\frac{\partial p}{\partial \omega} > 0$ (according to (2.12)). Because of the relative price effect p, which now is at work within the finance sector, the demand for risky assets is substituted by demand for safe assets if the relative price of services for securities rises. For low values of the skill premium, we are on the downward sloping branch of the $\tilde{\eta}$-curve so that income and substitution effect go in the same direction. In the upward sloping part of $\tilde{\eta}$, the negative substitution effect is opposed by a positive income effect so that the total effect of ω on Z_2^D depends on the relative importance of the two effects. Numerical simulation shows that the substitution effect is large if the price p_{z_2} is high and the income effect is stronger if subsistence expenditures are larger. For a high level of price p_{z_2} (based on (2.12) this means, for example, a low A_{z_2}) and low subsistence levels (such that $\bar{y}$ is close to zero) the substitution effect dominates. In this case $\frac{\partial Z_2^D}{\partial \omega} < 0$. However, for low levels of price p_{z_2} and large subsistence levels the income effect dominates. For this case, (B.25) and (B.26) give us the same picture as in Figure 2.4. Proposition 2.3 remains valid in both cases.

For Proposition 2.4, we have to write the excess demand function $Z_2^D - Z_2^S$ explicitly in terms of parameters. Using $W = b_L \bar{L} A_x \Gamma_x \omega^{-\alpha_x}(1+\omega k)$ and $p_{z_2} = \frac{A_x \Gamma_x}{A_{z_2} \Gamma_{z_2}} \omega^{\alpha_{z_2}-\alpha_x}$ in (B.26), we can rewrite the equilibrium condition $Z_2^D - Z_2^S = 0$ in the form:

$$\begin{aligned}
&\frac{\mu-\rho\frac{1+\frac{A_x\Gamma_x}{A_{z_2}\Gamma_{z_2}}\omega^{\alpha_{z_2}-\alpha_x}}{1+p_{z_1}}}{1-\rho\frac{1+\frac{A_x\Gamma_x}{A_{z_2}\Gamma_{z_2}}\omega^{\alpha_{z_2}-\alpha_x}}{1+p_{z_1}}} \frac{\delta}{1+\delta} \frac{\Gamma_x \omega^{-\alpha_x}(1+\omega k) - \frac{N}{b_L \bar{L} A_x}\bar{y}}{1+\frac{A_x\Gamma_x}{A_{z_2}\Gamma_{z_2}}\omega^{\alpha_{z_2}-\alpha_x}} - \frac{A_{z_2}}{A_x}\frac{\gamma_{z_2}^{\alpha_{z_2}}}{\gamma_{z_2}-\gamma_x}\omega^{-\alpha_{z_2}}(k\omega-\gamma_x) \\
&\equiv D\left[\omega \middle| \underset{?}{\frac{A_{z_2}}{A_x}}, \underset{?}{k}, \underset{+}{\frac{A_x b_L \bar{L}}{N}}, \underset{-}{\bar{y}}, \underset{+}{\mu}, \underset{-}{\rho}, \underset{+}{\delta}\right] = 0.
\end{aligned}$$

Hence, an increase of $\frac{A_x b_L \bar{L}}{N}$ always leads to a rise in the equilibrium skill premium. Under Assumption 2.2, this was only the case if present subsistence expenditure dominates

[8] The only thing that changes is that now we have $\frac{\bar{y}}{1+p_{z_2}}$ with $\bar{y}$ constant instead of $\frac{\bar{y}}{1+p_{z_1}} = \frac{\bar{e_0}}{1+p_{z_1}} - \frac{\bar{e_1}}{r}$. Thus, apart from subscript z_2 instead of $z(= z_1 = z_2)$ in the modified proof we have $\bar{y}$ instead of $\bar{e_0}$ and no negative term $-\frac{\bar{e_1}}{r}$.

futures subsistence requirements (Proposition 2.4). Moreover, a decline in subsistence requirements $\bar{y}$ has unambiguously a positive impact on the equilibrium skill premium - regardless of whether the decline in $\bar{y}$ is caused by a decline in $\bar{e_0}$ or $\bar{e_1}$.

In contrast to the benchmark analysis, the equilibrium skill premium is now also affected by changes in μ and ρ. Finally, a rise in δ has now an unambiguously positive effect on ω^*. (In the benchmark analysis the role of δ was ambiguous.) The following proposition summarizes the comparative-static effects on the equilibrium skill premium under Assumption 2'.

Proposition 1.5'. *If Assumption 2.2 is replaced by Assumption 2', then:*

a) For $\bar{y} > 0$, a rise in $\frac{A_x b_L \bar{L}}{N}$ (caused by uniform technical progress or education and biased progress) raises the equilibrium skill premium. A decline of total subsistence requirements $\bar{y}$ (wherever they come from) have the same effect.

b) Financial innovations (a rise in μ) or increased attractiveness of risky investments (a decline of ρ) raise the equilibrium skill premium. A lower discount on the future (a rise of δ) has the same effect. These effects also hold if $\bar{y} = 0$.

Proof. Main text. □

As a consequence of (B.24), Proposition 2.5 remains valid if applied to the structure between new finance on the one side and production cum traditional finance on the other side. We have

Proposition 1.6'. *At given $\frac{A_z}{A_x}$, k, any change in other exogenous fundamentals which raises the skill premium leads to structural change from production and traditional finance (X^+) towards new finance (Z_2).*

Proof. Equation (B.24). □

Finally, the ratio of value-added in financial subsector Z_2 to value-added in subsector Z_1 is as in (2.27)

$$\frac{p_{z_2} F}{p_{z_1} D} = \frac{s_f \bar{\eta}}{s_d \bar{\eta} + \frac{1+\delta}{\delta} \frac{\bar{e_1}}{r}} \frac{p_{z_2}}{1 + p_{z_2}} \frac{1 + p_{z_1}}{p_{z_1}}. \tag{B.27}$$

Since p_{z_1} and $\bar{y}$ are constant, $\frac{\partial \bar{w}}{\partial \omega} > 0$ immediately implies $\frac{\partial \bar{\eta}}{\partial \omega} > 0$. Hence, for $\bar{e}_1 > 0$, the income effect unambiguously leads to structural change from Z_1 to Z_2 if the skill premium rises. If $\bar{e}_1 = 0$, no such income effect is at work; yet the relative price effect remains. For the relative price effect, we only have to consider p_{z_2} because p_{z_1} is constant. Price p_{z_2} affects the value added structure within finance through two channels: On the one side, there is the direct effect shown explicitly in (B.27). Since $\frac{\partial p_{z_2}}{\partial \omega} > 0$, this channel tends to increase the share of new finance. On the other side, however, there is the negative substitution effect in the demand for financial services ($\frac{\partial s_f}{\partial p} < 0$ and $\frac{\partial s_d}{\partial p} > 0$) which drives

the sectoral structure within finance from Z_2 towards Z_1. Due to this ambiguous role of the relative price effect under the alternative specification, within structural change from Z_1 to Z_2 is more difficult to model than it was in the benchmark. For high levels of price p_{z_2} and low subsistence expenditures the substitution effect dominates. Then, the presented model cannot predict a co-movement of ω and the within structural change from Z_1 to Z_2. In the other case, however, Proposition (2.6) applies.

B.4 Data

Table B.1: Parameters survey years 1995-2009

Parameter	Data	Source	Description
$\bar{L}$	109m	CPS	# Low-skilled employees
$\bar{H}$	41.1m	CPS	# High-skilled employees
h^L	1755.6	CPS	Yearly hours of low-skilled
h^H	2025.3	CPS	Yearly hours of high-skilled
α_x	0.44	CPS	Output ela. of high-skilled in X
α_{z_1}	0.54	CPS	Output ela. of high-skilled in Z_1
α_{z_2}	0.79	CPS	Output ela. of high-skilled in Z_2
A_x	32.46	CPS	Technology level in X
PT_{65}	\$ 11,213	U.S. Bureau of the Census	Real poverty threshold <65
PT^{65}	\$ 10,080	U.S. Bureau of the Census	Real poverty threshold >65
$LEratio$	3.83	LE from World Bank	Old-age ratio
r^f	0.0151	Federal Reserve Bank of St.Louis	Real federal funds rate
A_{z_1}	141.93	Model calibration·A_x-growth	Technology level in Z_1
A_{z_2}	201.88	Model calibration·A_x-growth	Technology level in Z_2
δ	0.385	Model calibration	Discount rate
μ	0.740	Model calibration	Certainty measure

Notes: The table shows the averaged values for the time range of survey years $t \in \{1995, \ldots, 2009\}$. Averages of $\alpha_{j,t} = \frac{\kappa_{j,t}\omega_{j,t}}{1+\kappa_{j,t}\omega_{j,t}}$ with $\kappa_{j,t} = \frac{h^H_{j,t}\bar{H}_{j,t}}{h^L_{j,t}\bar{L}_{j,t}}$ and $\omega_{j,t} = \frac{w^H_{j,t}}{w^L_{j,t}}$, $j \in \{x, z_1, z_2\}$, $h^H_t = h^H_{x,t}$ and $h^L_t = h^L_{x,t}$. $A_{x,t} = \frac{w^L_{x,t}}{\Gamma_{x,t}\omega_{x,t}^{-\alpha_{x,t}}}$ with $\Gamma_{x,t} = {\alpha_{x,t}}^{\alpha_{x,t}}(1-\alpha_{x,t})^{1-\alpha_{x,t}}$. PT is the average, real poverty threshold of a two-people household (nominal values are adjusted by using the CPI-U adjustment factor to 1999 dollars (i.e., for the base survey year 2000) from CPS with PT_{65} denoting the relevant value for households younger than 65 and PT^{65} denoting the value relevant for older ones. $LEratio$ is the average ratio of working-time to retirement: $(65-20)/(LE_t - 65)$, where LE_t denotes life expectancy in year t; 65 is the retirement age and 20 is the assumed start of the working-life. r^f is the average, real effective federal funds rate (effective federal funds rate adjusted with the CPI-U adjustment factor from CPS). See bibliography for details on data sources.

C Appendix: Chapter 3

C.1 Derivations and proofs

C.1.1 Derivation of the investment allocation of benevolent local officials

Plugging the fiscal budget constraint into the objective function (3.9) and taking first order condition, we get

$$\underbrace{\left[\pi(a,I_2;\theta)\frac{\partial U}{\partial C}\Big|_{E=E^H}+(1-\pi(a,I_2;\theta))\,\frac{\partial U}{\partial C}\Big|_{E=E^L}\right]}_{\equiv A}\frac{\partial Y}{\partial I_1}(1-\tau)$$
$$-\underbrace{\left[\pi(a,I_2;\theta)\,(\bar{\varphi}-\varphi(\theta))\,\frac{\partial U}{\partial E}\Big|_{E=E^H}+(1-\pi(a,I_2;\theta))\,\bar{\varphi}\frac{\partial U}{\partial E}\Big|_{E=E^L}\right]}_{\equiv B}\frac{\partial Y}{\partial I_1}$$
$$+\frac{\partial \pi}{\partial I_2}\frac{\partial I_2}{\partial I_1}\left[U(C,E^H)-U(C,E^L)\right]=0 \tag{C.1}$$

The three lines in equation (C.1) represent the tradeoff of local officials in choosing the optimal investment I_1 (and I_2 correspondingly using the budget constraint): Marginal utility from consumption (A), marginal utility from environment (B), and the change in utility due to a marginal change in the probability of ending up in a good state, respectively.

Since $\frac{\partial I_2}{\partial I_1}=\frac{\partial Y}{\partial I_1}\tau-1$, we can divide (C.1) by $\frac{\partial Y}{\partial I_1}$ and have

$$(1-\tau)A=B-\underbrace{\frac{\partial \pi}{\partial I_2}\left(\tau-1\Big/\frac{\partial Y}{\partial I_1}\right)}_{\equiv X}\underbrace{\left[U(C,E^H)-U(C,E^L)\right]}_{\equiv Z}, \tag{C.2}$$

where X represents the marginal impact of an increase in I_1 on the probability of successful policy implementation, and Z is the gap in utility between H and L state.

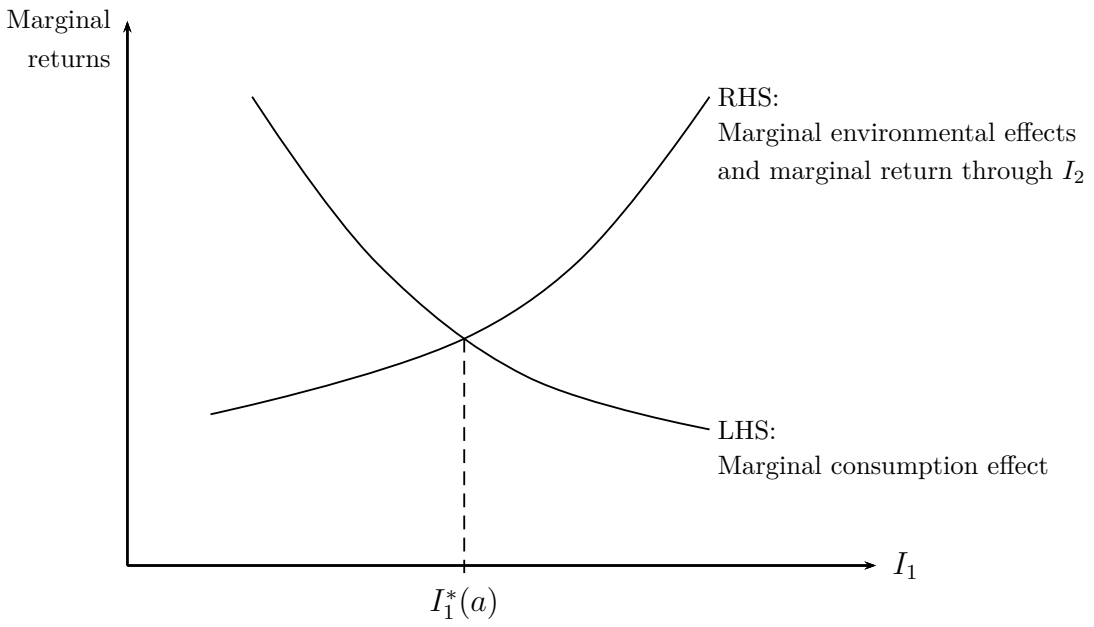

Figure C.1: Determination of socially optimal investment allocation

To see how the optimal investment is determined, we look at how the expressions on the LHS and on the RHS of (C.2) react to a change in I_1.

A decreases as I_1 increases, because output Y increases and drives down marginal utility from consumption. We illustrate this by a decreasing curve (i.e., LHS) in Figure C.1.

B increases as I_1 increases. As output Y increases with I_1, consumption increases and environmental quality decreases. This drives up the marginal utility from environment. The last term is less straightforward analytically. To see this more clearly, we take the partial derivative of the last term with respect to I_1. Namely,

$$\frac{\partial(XZ)}{\partial I_1} = \frac{\partial X}{\partial I_1} Z + \frac{\partial Z}{\partial I_1} X. \tag{C.3}$$

First, since utility is higher in a good state, $Z > 0$. taking derivative of X with respect to I_1, we have

$$\frac{\partial X}{\partial I_1} = 1 \Big/ \left(\frac{\partial Y}{\partial I_1}\right)^2 \left[\frac{\partial^2 \pi}{\partial I_2^2}\left(\tau \frac{\partial Y}{\partial I_1} - 1\right)^2 \frac{\partial Y}{\partial I_1} + \frac{\partial \pi}{\partial I_2}\frac{\partial^2 Y}{\partial I_1^2}\right] < 0, \tag{C.4}$$

because $\frac{\partial^2 \pi}{\partial I_2^2} < 0$ and $\frac{\partial^2 Y}{\partial I_1^2} < 0$. Therefore, $\frac{\partial X}{\partial I_1} Z < 0$. However, the sign of the second term

in (C.3) is unclear. Specifically,

$$\frac{\partial Z}{\partial I_1} = \underbrace{\left(\frac{\partial U}{\partial C}\Big|_{E^H} - \frac{\partial U}{\partial C}\Big|_{E^L}\right)}_{\equiv\Delta}(1-\tau)\frac{\partial Y}{\partial I_1} + \underbrace{\left((\bar{\varphi}-\varphi(\theta))\,\frac{\partial U}{\partial E}\Big|_{E^H} - \bar{\varphi}\frac{\partial U}{\partial E}\Big|_{E^L}\right)}_{\equiv\Lambda}\frac{\partial Y}{\partial I_1} \tag{C.5}$$

The marginal utility from consumption is higher when environmental quality is better, and thus $\Delta > 0$. However, the marginal utility from environment is higher when environmental quality is lower (i.e., $\frac{\partial U}{\partial E}\Big|_{E^H} < \frac{\partial U}{\partial E}\Big|_{E^L}$). Thus, $\Lambda < 0$. Therefore, when the impact on marginal utility from consumption is stronger, $\frac{\partial Z}{\partial I_1}$ is positive. Whereas if the impact on marginal utility from environment is stronger, $\frac{\partial Z}{\partial I_1}$ is negative. Furthermore, X is positive if an increase in I_1 increases I_2, which is the case if the increase in fiscal budget (i.e., $\tau Y(a, I_1)$) due to higher production-related infrastructure investment I_1 is more than the increase in the investment, and vice versa. Therefore, in the end the sign of the last term in (C.3) depends on the relative magnitude of the marginal utility from consumption and from environment, and whether an increase in I_1 would increase I_2.

Given the specification of the functional forms and parameter values, our numerical results indicate that as I_1 increases, XZ decreases (i.e., the negative sign of the first term in (C.3) dominates). Together with the fact that the marginal utility from environment, B, increases as I_1 increases, we know that the RHS increases as I_1 increases. This is illustrated by the increasing curve in Figure C.1 (i.e., RHS). The investment allocation under social optimum, $I^*(a)$, is determined by the intersection of the two curves in Figure C.1.

C.1.2 Proof of Proposition 3.1

Before proving the proposition, first notice that to achieve a given level of output, Y_0, there exists a minimal amount of investment in production-related infrastructure, I_1, for each level of ability, $a \in [a_{min}, a_{max}]$. Formally,

$$Y(a, I_1) = Y_0,$$

which gives $I_1 = I(\underset{-}{a}, \underset{+}{Y_0})$. The sign below the variables indicates the sign of the corresponding partial derivatives. Therefore, the output thresholds, $\left\{\underline{Y}^H, \underline{Y}^L, \bar{Y}^L, \bar{Y}^H\right\}$, discussed in Lemma 3.1 case iii) define corresponding thresholds of investment, $\left\{\underline{I}^H, \underline{I}^L, \bar{I}^L, \bar{I}^H\right\}$. We plot the four thresholds of investment and the maximal feasible investment in Figure C.2. The region below I_1^{max} represents the levels of investment I_1 that are feasible for officials of the respective ability levels. Any level above I_1^{max} is infeasible. By definition, each curve $\{\underline{I}_1^s\}_{s\in\{H,L\}}$ indicates the minimal amount of investment required to achieve

the corresponding production level.

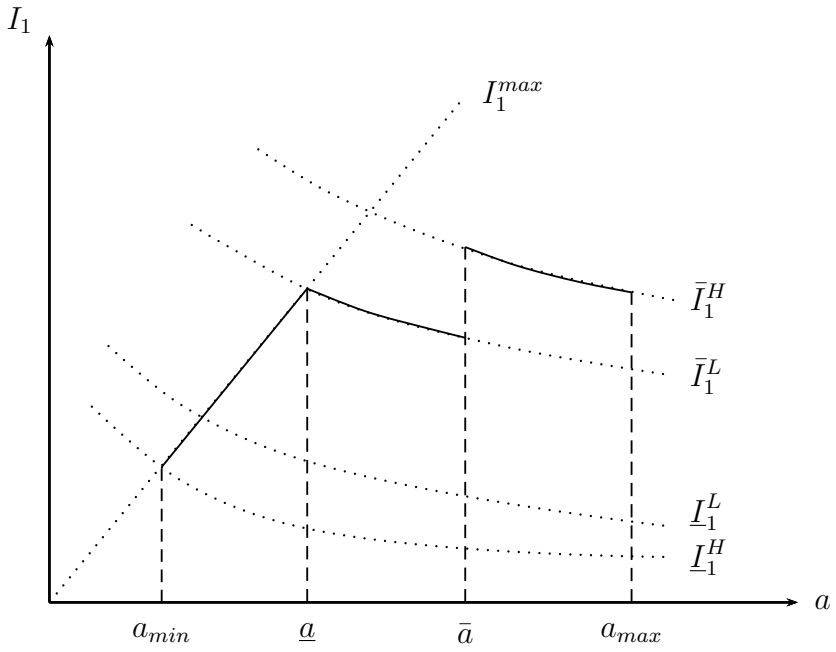

Figure C.2: Illustration of investment thresholds and maximal feasible investment

The minimal ability a_{min} is chosen such that the maximal feasible production is above $\underline{Y}^H$ for all officials. In other words, it is possible to choose an investment allocation so that households do not protest at least in high state. Therefore, to assume $a > a_{min}$ is equivalent to say that for all officials in the economy, they may stay in office or being promoted with positive probability.

Proof of Proposition 3.1. For officials with ability $a \in [a_{min}, \underline{a})$, their maximal feasible investment is below the $\bar{I}_1^L$ curve. This means that these officials will never be protested against due to overproduction, because the production thresholds $\bar{Y}^s, s \in \{H, L\}$ are not achievable. Therefore, plugging the expression of $\hat{Y}$ in (3.15), we have for the officials' maximization problem

$$\pi^D \underline{A} + (1 - \pi^D) \left[\bar{A} - \Phi(Y^* - Y)(\bar{A} - M) \right],$$

with

$$\pi^D = \pi(a, I_2; \theta) \mathbb{1}_{\left\{Y \leq \underline{Y}^H\right\}} + (1 - \pi(a, I_2; \theta)) \mathbb{1}_{\left\{Y \leq \underline{Y}^L\right\}},$$

and $\Phi(\cdot)$ be the distribution function of $\epsilon \sim N(0, \sigma^2)$.

Notice that $\Phi(Y^* - Y)$ increases in Y, the higher the output, the more likely the local official is promoted. In addition, as output level increases, the probability of household

protest due to low consumption is lower (i.e., π^D decreases in Y). Therefore, the optimal choice of the local officials with ability in this interval is to invest all resources into production to boost output level. This is illustrated by the line segment of I_1^{max} between $[a_{min}, \underline{a})$.

Now we consider officials with ability $a \in [\underline{a}, a_{max}]$.

We claim that investment levels above $\bar{I}^H$ can never be an optimal solution. At this investment level local output is above $\bar{Y}^H$. Households protest due to bad environment in both states. In other words, local officials will be demoted for sure. Officials who are able to produce $\bar{Y}^H$ can always decrease the level of production and thus avoid households' protest by invest more resources into environment-related infrastructure. Therefore, $\bar{I_1}^H$-curve gives the upper bound of the optimal investment allocation.

In addition, investment levels below $\bar{I}^L$ are also not optimal. Since local officials will not induce household protest at production below $\bar{Y}^L$, they have an incentive to produce at least $\bar{Y}^L$, when it is achievable. In other words, for local officials with ability $a \in [\underline{a}, a_{max}]$, the optimal allocation problem is to determine whether to produce above $\bar{Y}^L$. Furthermore, if local officials produce above $\bar{Y}^L$ but below $\bar{Y}^H$, household protest for sure in a bad state, and not in a good state. Therefore, the probability of demotion is equal to the probability of policy failure, $\pi^D = 1 - \pi(a, I_2; \theta)$. And the maximization problem is given by

$$\begin{aligned} \max_{\{I_1, I_2\}} \quad & U^{LO}(I_1, I_2, a, \theta), \\ s.t. \quad & U^{LO}(I_1, I_2, a, \theta) \geq \Gamma(\bar{Y}^L), \end{aligned} \tag{C.6}$$

where $U^{LO}(I_1, I_2, a, \theta)$ is defined in (3.15), and $\Gamma(\bar{Y}^L)$ is local officials' reward when producing at output level $\bar{Y}^L$. □

C.1.3 Ratio of investment allocation

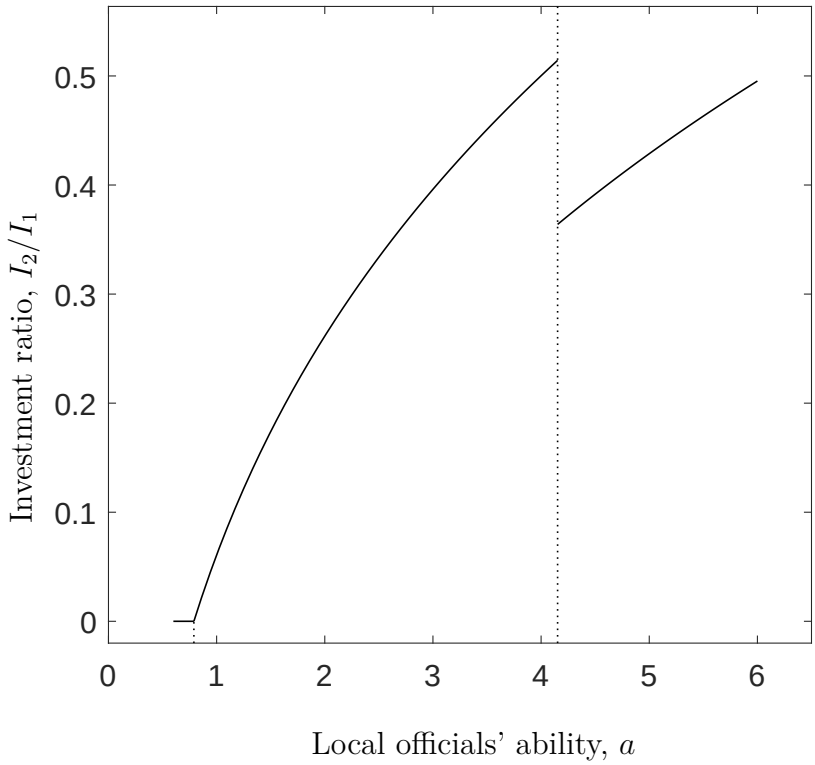

Figure C.3: Investment ratio, I_2/I_1. The two vertical dotted lines indicate the lower and upper threshold of local officials' ability, $\underline{a}$ and $\bar{a}$, respectively.

Part IV

Bibliography

Bibliography

Acemoglu, D. (2002). Directed technical change. *Review of Economic Studies 69(4)*, 781–809.

Acemoglu, D. and F. Zilibotti (1997). Was Promotheus unbound by chance? Risk, diversification, and growth. *Journal of Political Economy 105(4)*, 709–751.

Achury, C., S. Hubar, and C. Koulovatianos (2012). Saving rates and portfolio choice with subsistence consumption. *Review of Economic Dynamics 15*(1), 108–126.

Admati, A. and M. Hellwig (2013). *The bankers' new clothes: What's wrong with banking and what to do about it*. Princeton: Princeton University Press.

Albuquerque, R. and H. A. Hopenhayn (2004). Optimal lending contracts and firm dynamics. *Review of Economic Studies 71(2)*, 285–315.

Alesina, A. and G. Tabellini (2007). Bureaucrats or politicians? Part I: A single policy task. *American Economic Review 97*(1), 169–179.

Angelini, P. and A. Generale (2008). On the evolution of firm size distributions. *American Economic Review 98(1)*, 426–438.

Asian Development Bank (2014). *Local public finance management in the People's Republic of China: Challenges and opportunities*. Mandaluyong City, Philippines: Asian Development Bank.

Atkeson, A. and R. E. J. Lucas (1992). On efficient distribution with private information. *Review of Economic Studies 59(3)*, 427–435.

Atkeson, A. and R. E. J. Lucas (1995). Efficiency and equality in a simple model of efficient unemployment insurance. *Journal of Economic Theory 66(1)*, 64–88.

Atkinson, A. B. (1997). Bringing income distribution in from the cold. *The Economic Journal 107(441)*, 297–321.

Beck, T., B. Büyükkarabacak, F. K. Rioja, and N. T. Valev (2012). Who gets the credit? and does it matter? household vs. firm lending across countries. *The B.E. Journal of Macroeconomics 12*(1), 1–44. Article 2.

Berger, A. N. and G. F. Udell (2002). Small business credit availability and relationship lending: The importance of bank organisational structure. *The Economic Journal 112(477)*, F32–F53.

Besley, T. and A. Case (1995). Does electoral accountability affect economic policy choices? evidence from gubernatorial term limits. *Quarterly Journal of Economics*, 769–798.

Besley, T. J. and R. Burgess (2002). The political economy of government responsiveness: Theory and evidence from india. *Quarterly Journal of Economics*, 1415–1451.

Boppart, T. (2014). Structural change and the Kaldor facts in a growth model with relative price effects and non-Gorman preferences. *Econometrica 107(441)*, 297–321.

Boppart, T. (2015). To which extent is the rise in the skill premium explained by an income effect? *mimeo*.

Cabral, L. M. and J. Mata (2003). On the evolution of the firm size distribution: Facts and theory. *American Economic Review 93(4)*, 1075–1090.

Capelle-Blancard, G. and C. Labonne (2011). More bankers, more growth? evidence from OECD countries. *CEPII Document de Travail 2011-22*.

Cecchetti, S. G. and E. Kharroubi (2015). Why does financial sector growth crowd out real economic growth? *BIS Working Papers 490*.

Célérier, C. and B. Vallée (2014). The motives for financial complexity: An empirical investigation. *HEC Paris Research Paper No. FIN-2013-1013*.

Célérier, C. and B. Vallée (2016). Returns to talent and the finance wage premium.

mimeo.

Chang, L., W. Li, and X. Lu (2015). Government engagement, environmental policy, and environmental performance: evidence from the most polluting chinese listed firms. *Business Strategy and the Environment 24*(1), 1–19.

Clementi, G. L. and H. A. Hopenhayn (2006). A theory of financing constraints and firm dynamics. *Quaterly Journal of Economics 121(1)*, 229–265.

DeMarzo, P. M. and M. J. Fishman (2007). Agency and optimal investment dynamics. *Review of Financial Studies 20(1)*, 151–188.

Dyrda, S. (2016). Fluctuations in uncertainty, efficient borrowing constraints and firm dynamics. *mimeo*.

Edin, M. (2003). State capacity and local agent control in China: CCP cadre management from a township perspective. *The China Quarterly 173*, 35–52.

Evans, D. S. (1987). The relationship between firm growth, size, and age: Estimates for 100 manufacturing industries. *Journal of Industrial Economics 35(4)*, 567–581.

Falkinger, J. (2014). In search of economic reality under the veil of financial markets. *Working paper series / Department of Economics 154*.

Federal Reserve Bank of St.Louis (accessed 12.08.2015). Effective federal funds rate. https://research.stlouisfed.org/fred2/series/DFF.

Föllmi, R. and J. Zweimüller (2008). Structural change, Engel's consumption cycles and Kaldor's facts of economic growth. *Journal of Monetary Economics 55(7)*, 1317–1328.

Fredriksson, P. G. and J. R. Wollscheid (2014). Political institutions, political careers and environmental policy. *Kyklos 67*(1), 54–73.

Gennaioli, N., A. Shleifer, and R. W. Vishny (2014). Finance and the preservation of wealth. *Quarterly Journal of Economics 129(3)*, 1221–1254.

Green, E. J. (1987). Lending and the smoothing of uninsurable income. In E. C. Prescott and N. Wallace (Eds.), *Contractual Arrangements for Intertemporal Trade*. University

of Minnesota Press, Minneapolis.

Greenwood, J. and B. Jovanovic (1990). Financial development, growth, and the distribution of income. *Journal of Political Economy 98(5)*, 1076–1107.

Greenwood, J. and D. Scharfstein (2013). The Growth of Finance. *Journal of Economic Perspectives 27(2)*, 3–28.

Gross, T. and S. Verani (2013). Financing constraints, firm dynamics, and international trade. *Finance and Economics Discussion Series 2013-02*. Washington: Board of Governors of the Federal Reserve System.

Gründler, K. and J. Weitzel (2012). The financial sector and economic growth in a panel of countries. *Wirtschaftswissenschaftliche Beiträge des Lehrstuhls für Volkswirtschaftslehre, insbes. Wirtschaftsordnung und Sozialpolitik, Universität Würzburg Prof. Dr. Norbert Berthold 123*.

Hall, B. H. (1987). The relationship between firm size and firm growth in the U.S. manufacturing sector. *Journal of Industrial Economics 35(4)*, 583–606.

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *Review of Economic Studies 66*(1), 169–182.

Jia, R. (2014). Pollution for promotion. *Mimeo*. UCSD.

Jin, H., Y. Qian, and B. R. Weingast (2005). Regional decentralization and fiscal incentives: Federalism, Chinese style. *Journal of Public Economics 89*(9), 1719–1742.

Judd, K. L. (1998). *Numerical Methods in Economics*. Cambridge, Massachusetts: MIT Press.

King, M., S. Ruggles, J. T. Alexander, S. Flood, K. Genadek, M. B. Schroeder, B. Trampe, and R. Vick (2010). *Integrated Public Use Microdata Series, Current Population Survey: Version 3.0. [Machine-readable database]*. Minneapolis: University of Minnesota.

Kneer, C. (2013). The Absorption of Talent into Finance: Evidence from U.S. Banking Deregulation. *DNB Working Paper 391*.

Law, S. H. and N. Singh (2014). Does too much finance harm economic growth. *Journal of Banking and Finance 41*, 36–44.

Levine, R. (2005). Finance and growth: Theory and evidence. *Handbook of Economic Growth 1A*, 855–934.

Li, D. D. (1998). Changing incentives of the Chinese bureaucracy. *American Economic Review*, 393–397.

Li, H. and L.-A. Zhou (2005). Political turnover and economic performance: the incentive role of personnel control in China. *Journal of Public Economics 89*(9), 1743–1762.

List, J. A. and D. M. Sturm (2006). How elections matter: Theory and evidence from environmental policy. *Quarterly Journal of Economics*, 1249–1281.

Ljungqvist, L. and T. J. Sargent (2000). *Recursive macroeconomic theory*. Cambridge, Massachusetts: MIT Press.

Machin, S. and J. Van Reenen (1998). Technology and changes in skill structure: Evidence from seven OECD countries. *Quarterly Journal of Economics 113(4)*, 1215–1244.

Maskin, E., Y. Qian, and C. Xu (2000). Incentives, information, and organizational form. *Review of Economic Studies 67*(2), 359–378.

OECD Data (accessed 29.06.2015). Employment: Self-employment rate. https://data.oecd.org/emp/self-employment-rate.htm.

Philippon, T. (2012). Notes on equilibrium financial intermediation. *mimeo NYU*.

Philippon, T. (2015). Has the US Finance Industry Become Less Efficient? On the Theory and Measurement of Financial Intermediation. *American Economic Review 105(4)*, 1408–1438.

Philippon, T. and A. Reshef (2007). Skill biased financial development: Education, wages and occupations in the U.S. financial sector. *NBER Working Paper 13437*.

Philippon, T. and A. Reshef (2012). Wages and human capital in the U.S. financial industry: 1909-2006. *Quaterly Journal of Economics 127(4)*, 1551–1609.

Philippon, T. and A. Reshef (2013). An international look at the growth of modern finance. *Journal of Economic Perspectives 27(2)*, 73–96.

Piketty, T. (2014). *Capital in the twenty-first century*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Piketty, T. and E. Saez (2003). Income inequality in the United States, 1913-1998. *Quarterly Journal of Economics 118(1)*, 1–39.

Quadrini, V. (2004). Investment and liquidation in renegotiation-proof contracts with moral hazard. *Journal of Monetary Economics 51(4)*, 713–751.

Radner, R. (1985). Repeated principal-agent games with discounting. *Econometrica 53(5)*, 1173–1198.

Rochlitz, M., V. P. Kulpina, T. F. Remington, and A. A. Yakovlev (2014). Performance incentives and economic growth: Regional officials in Russia and China. *Higher School of Economics Research Paper No. WP BRP 18*.

Rogerson, W. P. (1985). Repeated moral hazard. *Econometrica 53(1)*, 69–76.

Rousseau, P. L. and P. Wachtel (2011). What is happening to the impact of financial deepening on economic growth? *Economic Inquiry 49*(1), 276–288.

Saich, T. (2012). The quality of governance in China: The citizens' view. *HKS Faculty Research Working Paper Series, RWP12-051*.

Smith, A. A. J. and C. Wang (2006). Dynamic credit relationships in general equilibrium. *Journal of Monetary Economics 53(4)*, 847–877.

Spear, S. E. and S. Srivastava (1987). On repeated moral hazard with discounting. *Review of Economic Studies, 54(4)*, 599–617.

Studer, S. (2015). An equilibrium model with diversification-seeking households, competing banks and (non-)correlated financial innovations. *mimeo*.

Su, F., R. Tao, L. Xi, and M. Li (2012). Local officials' incentives and China's economic growth: Tournament thesis reexamined and alternative explanatory framework. *China*

& World Economy 20(4), 1–18.

Suellow, T. (2015). The skill-intensity of financial service consumption: An input-output analysis. *mimeo*.

The Economist (2015). As safe as houses. *The Economist 2015, January 31*, 60.

Thomas, J. and T. Worrall (1990). Income fluctuation and asymmetric information: An example of a reapeated principal-agent problem. *Journal of Economic Theory 51(2)*, 367–390.

Townsend, R. M. (1982). Optimal multiperiod contracts and the gain from enduring relationships under private information. *Journal of Political Economy 90(6)*, 1166–1186.

Tsui, K. and Y. Wang (2004). Between separate stoves and a single menu: fiscal decentralization in China. *The China Quarterly 177*, 71–90.

U.S. Bureau of the Census (accessed 12.08.2015). Average Poverty Thresholds. https://www.census.gov/hhes/www/poverty/data/historical/index.html.

Verani, S. (2015). Aggregate consequences of dynamic credit relationships. *Finance and Economics Discussion Series 2015-063*. Washington: Board of Governors of the Federal Reserve System, http://dx.doi.org/10.17016/FEDS.2015.063.

Wang, Q. (2010, May). China's environmental civilian activism. *Science 328*, 824.

World Bank (accessed 04.11.2015b). World Development Indicators: Life expectancy at birth, total (years). http://data.worldbank.org/indicator/SP.DYN.LE00.IN.

World Bank (accessed 05.11.2015a). World Development Indicators: Gross savings (percentage of GDP). http://data.worldbank.org/indicator/NY.GNS.ICTR.ZS.

Wu, J., Y. Deng, J. Huang, R. Morck, and B. Yeung (2013). Incentives and outcomes: China's environmental policy. *Working Paper 18754*. National Bureau of Economic Research.

Xu, C. (2011). The fundamental institutions of China's reforms and development. *Journal*

of Economic Literature, 1076–1151.

Zheng, S. and M. E. Kahn (2013). Understanding China's urban pollution dynamics. *Journal of Economic Literature 51*(3), 731–772.

Part V

Curriculum Vitae

Curriculum Vitae

Personal details

Name:	Yingnan Zhao
Date of Birth:	14 June, 1988
Place of Birth:	Fuxin, China
Nationality:	Chinese

Education

09/2011 – 09/2016	PhD studies at the Zurich Graduate School of Economics (Fast track) University of Zurich, Switzerland Supervisor: Prof. Dr. Dr. Josef Falkinger
09/2011 – 04/2013	Master of Science in Economics University of Zurich, Switzerland
09/2006 – 07/2011	Bachelor of Science in Mathematics Fudan University, China
09/2010 – 01/2011	Exchange Semester University of Zurich, Switzerland

Professional experience

09/2016 – 08/2017	Post-doctoral research fellow at the Department of Economics, University of Zurich
09/2011 – 08/2016	Research and teaching assistant at the Department of Economics, University of Zurich